A New Method for the Analysis of Pharmaceutical Plant Cleaning strategies

A thesis submitted by

Wendy Adele Carr

For the degree of Engineering Doctorate

School of Chemical Engineering and Advanced Materials

Newcastle University

May 2017

Abstract

Plant cleaning in the pharmaceutical industry is an undervalued but critical stage of processing. Cleaning with solvents or other cleaning agents is often the only method capable of removing residual particles.

In a typical multipurpose pharmaceutical plant, cleaning challenges can cost companies millions of pounds' as they must clean plant equipment effectively to satisfy regulatory constraints. Failure to do this right first time can result in missed processing schedules often with financial consequences. Furthermore, cleaning is often only considered once the process chemistry has been optimised.

The research presented in this thesis describes a new approach to understanding the science behind cleaning using multivariate data analysis, principal component analysis (PCA). This approach utilises the fact that cleaning agent selection can be determined based on the identification of chemical functional groups and physiochemical properties of pharmaceutical products. Using PCA, a set of products were identified which could potentially be cleaned utilising the same approach. This means that the selection of a cleaning agent can be determined for other products with the same chemical functional groups and physiochemical properties.

Adopting this methodology helps decide if the cleaning agent used is appropriate to the process chemistry and therefore cleaning can be carried out right first time and as cost effectively as possible.

The findings from this research were developed into a tool and used to support the design of manufacturing processes taking cleaning into account from early stages of development thereby saving time and money during the processing stages. Ultimately, the tool will be incorporated into a suite of original and adapted Britest Ltd tools, entitled Fundamental Understanding of Science and Engineering (FUSE) used to identify, understand and provide solutions for cleaning challenges. The tool was applied to industrial case studies to assess its potential.

Dedication

This research is dedicated to my husband Vincent, and my mother Margaret Horler.

Acknowledgements

This research would not have been possible without the help and support of my industrial supervisor Mark Talford and I am very grateful to him and everyone in the Britest team for their support. I also gratefully acknowledge the support and guidance of my academic supervisors Elaine Martin and Moritz von Stosch.

Finally, I would not have been able to carry out this research without the help and support of the Britest Ltd Industrial members who allowed me access to their manufacturing sites and gave me information on cleaning and equipment use on numerous occasions.

Table of Contents

Table of Contents

Table of Contents

Terminology:

Acronym	Explanation
API	Active Pharmaceutical Ingredient
ASOG	Analytical Solutions Of Groups
BGIT	Benson Group Increment Theory
CBA	Cost Benefit Analysis
CIP	Cleaning in Place
COSMO	Conductor like Screen Model
CPP	Critical Process Parameters
CQA	Critical Quality Attributes
DFA	Driving Force Analysis
DuDEs	Duty Definition and Equipment Specification
EFPIA	European Federation of Pharmaceutical Industries and Associations
EMIC	Electromagnetic ion cyclotron
FDA	Food and Drugs Administration
FUSE	Fundamental Understanding of Science and Engineering
GMP	Good Manufacturing Practice
HACCP	Hazard Analysis and Critical Control Points
HPLC	High Performance Liquid Chromatography
IBC	Intermediate Bulk Container
ICH	International Conference of Harmonisation
IChemE	Institution of Chemical Engineers
IMS	Ion Mobility Spectronomy
ISA	Initial Screening Analysis
MACO	Maximum Allowed Carry Over limit
MS	Mass Spectronomy
NME	New Molecular Entity
NMR	Nuclear Magnetic Resonance
PDD	Process Definition Diagram
PDCD	Process Definition Cleaning Diagram
PDDP	Principal Direction Divisive Partitioning
ppm	Part Per Million
PrISM	Process Information Summary Map
PTFE	Polytetrafluorethylene

Acronym	Explanation
PVC	Poly vinyl chloride
QSPR	Quantitative Structure Activity Relationship Modelling
R and D	Research and Development
RC	Rich Cartoon
RFT	Right First Time
RP	Rich Picture
RQ	Research Question
RSC	Royal Society of Chemistry
TACT	Time, Action, Concentration and Temperature
TM	Transformation Map
UNIFAC	Universal Functional Activity Coefficient Model
WPD	Whole Process Design
WPU	Whole Process Understanding
ZEAL	Zero Emissions through Advanced cLeaning

List of Figures:

List of Tables:

List of Equations:

Chapter 1. Thesis Motivation and Overview

1.1 Thesis Motivation

The European Federation of Pharmaceutical Industries and Associations (EFPIA) issuing key data for 2014 estimated that pharmaceutical production, research and development expenditure and sales have steadily risen since 1990, giving it the status as one of Europe's highest performing advanced technology segments (EFPIA, 2014). In order to achieve this status companies in the sector are required to overcome a significant number of challenges (Figure 1-1), not least of which are global and economic conditions which saw the market dip in 2008 – 2010, regulatory requirements and the threat of new market entrants. In addition trends such as a decline in pharmaceutical industry innovation, patent expiration and mergers have resulted in low research and development (R&D) productivity (Comanor and Scherer, 2012). CMR International (2013) stated the outlook for global R&D productivity is rather bleak but their figures show a number of encouraging trends in the pharmaceutical sector. These include an above average number of New Molecular Entity (NME) first world introductions in 2012 (Figure 1-2) and a reduction in overall development time from 15 years to a new average of 12 years (product to market). In addition, healthcare reforms in the two biggest pharmaceutical market leaders China and America indicate that spending on pharmaceuticals is increasing (Daemmrich and Mohanty 2014).

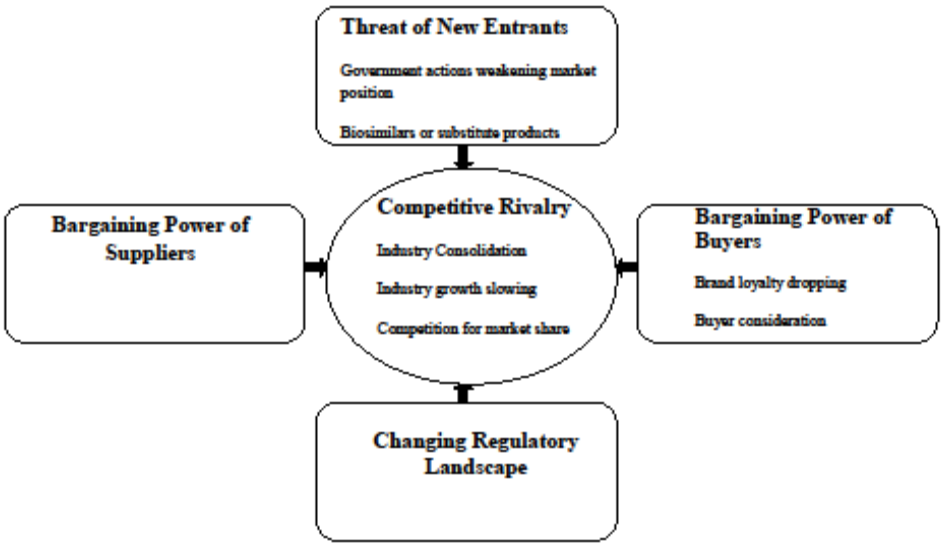


Figure 1-1 Competitive pressures and uncertainties. Adapted from Lainez, Schaefer and Reklatis, 2012.

In order to meet sector market challenges (Figure 1-1) including environmental issues and reducing R&D costs, R&D productivity needs to be addressed. It is considered that this challenge can only be addressed by proposing specific strategies (Paul, 2010). Research indicates that in order to optimise pharmaceutical and chemical processes, satisfy regulatory bodies and continue to manufacture effectively, the appropriate in depth process knowledge is critical.

Development of a strategy is common with respect to product development and this involves understanding pharmaceutical quality by design (Juran, 1992 and Yu, 2008). This has never been more pertinent than in the current economic and political climate. Global market conditions and the rise of personalised medicines require pharmaceutical and fine chemical companies to continuously improve processes and strive to reduce costs associated with all aspects of manufacturing. There are several challenges and business drivers associated with this including those identified by Khanna (2012) stating "low productivity, rising R&D costs, dissipating proprietary products and dwindling pipelines are driving the pharmaceutical industry to unprecedented challenges and scrutiny". According to research conducted by CMR in 2013 globally, pharmaceutical research and development is currently at a 16 year high with 39 New Drug Approvals and Biologics License Application approvals in 2012 (CMR, 2013). (Figure 1-2).

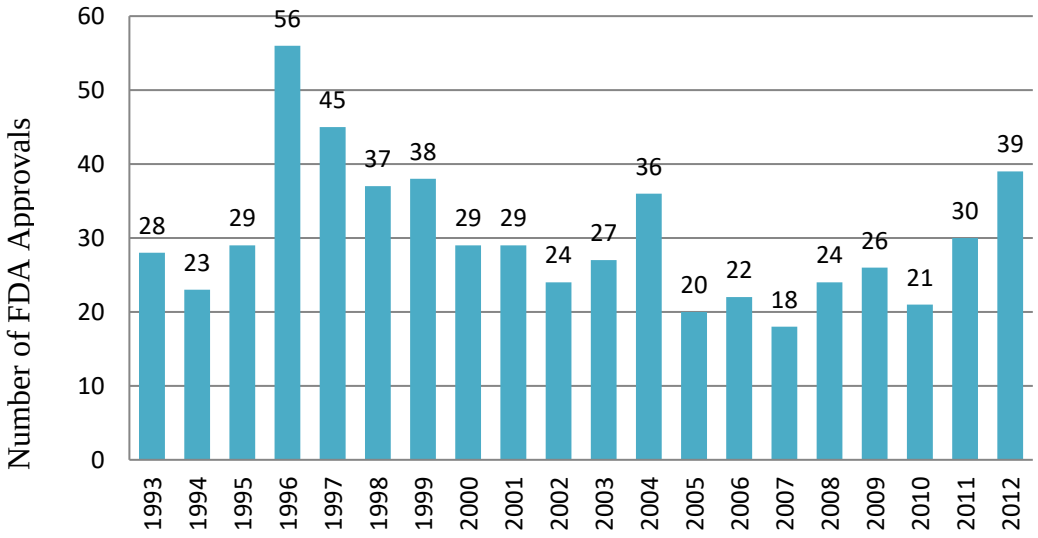


Figure 1-2 Number of Federal Drug Agency Approvals (FDA) per year between 1993 and 2012 (CMR, 2013)

The fine chemical industry is experiencing challenges of its own. Sustainable processes are sought and green chemistry has become a common phrase alongside safety, reducing cost, increasing quality and quantity. It is considered by many chemical bodies and associations

that the UK government needs to do more to support the fine chemical industry. In 2013 five groups formed an alliance to raise awareness and ask for more support. These include the Institution of Chemical Engineers (IChemE) and the Royal Society of Chemistry (RSC, 2013).

If the pharmaceutical industry and chemical industries are to operate successfully, they require process schedules to be efficient, product quality to be superior and optimised product batch size. This is important given that there are many stages involved in a drug reaching the market (Figure 1-3). It is important to consider that a potential new drug may fail at any stage of its development and the success rate for new Food and Drugs Administration (FDA) approvals per year is low (Figure 1-2) compared with the amount of potential drugs discovered (Figure1-3).

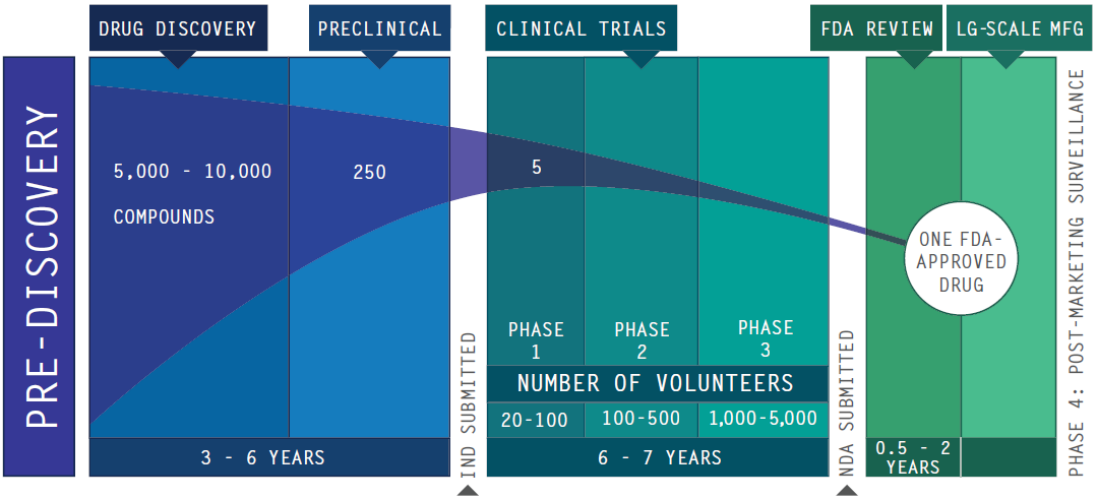


Figure 1-3 Drug Development Time (Hodgett, 2013)

In order to facilitate products through the development pipeline the consideration of Whole Process Design (WPD) is increasingly important. This is especially true in high value product manufacture. WPD assists the design of new products and helps determine which products are suitable to develop and designs processes to optimise criteria such as quality and quantity.

There are a number of design consultancies which aim to optimise and improve processes in various sectors utilising WPD. These companies include R B plant construction LTD (R B Plant, 2016), a company which provide engineering consultancy, design and construction management services to the process industries, and Allen Associates, who deliver whole process design for chemical and process engineering solutions, and have clients including ICI,

BARR and Drambuie (Allen Associates, 2016). Each company has a number of tools which enable WPD and Whole Process Understanding (WPU). This research is sponsored by Britest Limited and therefore the tools examined, used and developed are Britest Limited focussed for obvious reasons.

Britest Limited, a not for profit organisation, has developed a process understanding-based approach to WPD that enables organisations to develop ideas and address challenges in the pharmaceutical and chemistry industries. Britest members include industry and academia and ideas are shared and developed through collaboration. Britest develops tools and methodologies to address the identified challenges and provide solutions for industrial partners. The Britest tools and methodologies draw upon fundamental Whole Process Understanding (WPU) (Figure 1-4) to provide solutions. Implementing WPU allows a flexible approach where companies can develop techniques to enrich understanding of processes and drive innovation. Whole Process Understanding can be effectively used to influence WPD by considering the process as a whole, not as stages or steps. Redesign and optimisation is carried out with regard to the effect on the whole process from raw material entry to final formulation. The use of Britest's innovative techniques are estimated to have saved Britest industrial members in excess of £600 million between 2001 and 2012 (Britest Ltd, 2014).

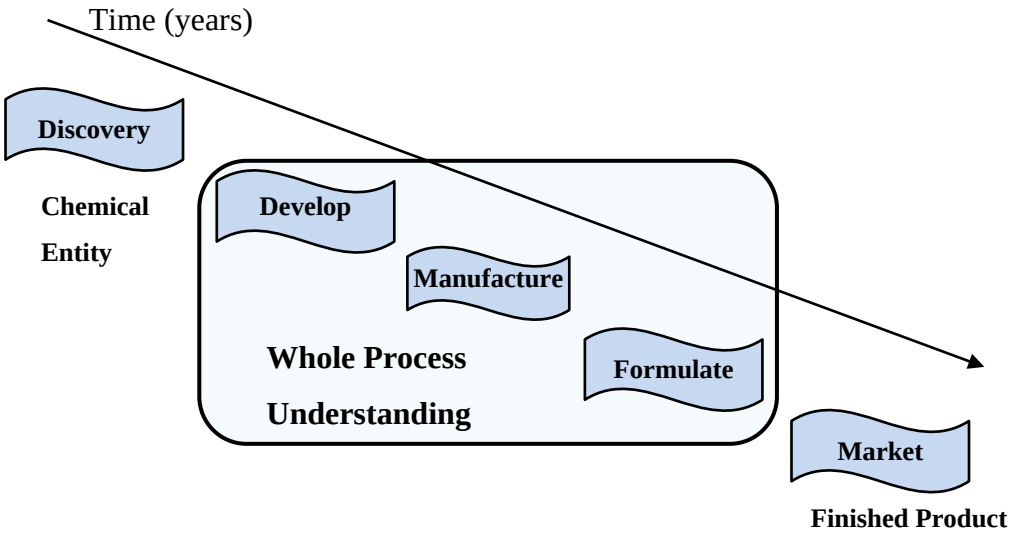


Figure 1-4 Whole Process Understanding (Adapted from Britest b, 2014)

The impact of Britest's approach to WPD is therefore significant, and can help to address specific industrial challenges, for example, deciding which raw materials to use in a process to prevent unwanted side reactions, or determining the phases of reactants in a vessel.

Methodologies and tools used by Britest lead to new innovation, but in order to achieve this in-depth process knowledge is required. This involves significant input from people who know and understand their products and processes. More importantly, it requires a multidisciplinary approach to achieve a fundamental understanding of science and engineering involved.

However, not all stages of a manufacturing process are considered or understood, even if a multidisciplinary approach is taken. Due to complexity some challenges take considerable effort to begin to understand. One of these challenges is the fundamental understanding of the science behind plant cleaning. It is considered that this is one of the most neglected and misunderstood areas of pharmaceutical and fine chemical production. Poor or ineffective cleaning has an effect on production time and scheduling leading to financial implications. There is also an effect on operator safety and environmental consideration. In well thought-out processes it is essential that information on cleaning is considered as all avenues of cost saving must be explored in R&D and process development in the pharmaceutical and fine chemical industries.

The current approaches to developing cleaning protocols are not perfect. It is often the case that cleaning protocols are only considered following process development, not as part of it. Cleaning protocols are therefore suboptimal and this has an impact on production costs, it increases the amount of plant and equipment downtime and leads to lost opportunity costs.

As stated Whole Process Design is essential in pharmaceutical and fine chemical manufacture, but how can Whole Process Understanding be considered complete without a fundamental knowledge of plant cleaning? Therefore, how can costs reduce without fundamental understanding of process plant cleaning? Further work must be carried out in order to address this gap in knowledge.

1.2 Aims and Objectives

The primary aim of this thesis is to develop an understanding of the fundamental science behind process plant cleaning. Key objectives include -

1. Understanding the current industrial requirements and limitations of cleaning
2. Production of a tool or set of tools to aid decision making for solvent/cleaning agent choice
3. Ensure the tool can be used early in the process design stage of manufacturing

4. Incorporate a whole process design methodology to plant cleaning by the inclusion of engineering and process considerations that affect cleaning

The creation of a suitable tool or the adaption of existing tools to aid cleaning agent/solvent choice will be used by Britest Limited to respond to the challenges associated with cleaning, for both the pharmaceutical industry and fine chemical sectors.

1.3 Research Questions

The main issue which drives the research in this thesis is:

RQ1: What would be the best way to increase the fundamental understanding of the science behind cleaning linked to solvent and cleaning agent use?

There are several research questions which underpin this question. These are:

RQ2: What is meant by the term "fundamental science" in relation to process plant cleaning?

RQ3: What are the main challenges associated with process plant cleaning for Britest members?

RQ4: What common methods do Britest members utilise to clean their process equipment?

RQ5: Which cleaning challenges are associated with plant engineering or choice of cleaning agent?

RQ6: What information is currently available in literature to help understand cleaning challenges?

RQ7: Which Britest tools and methodologies would be the best to examine cleaning and what adaptations would they require?

RQ8: How can process plant cleaning be demonstrated after the application of the knowledge gained in this thesis?

It is considered that the answers to these questions can be addressed by a Britest member survey, Britest member site visits and interviews, the literature review, an examination of Britest tools, methodologies and also industrial data collection.

During the course of this research further questions will be taken into account based on an appreciation of the knowledge and data accumulated.

1.4 Industrial Relationship

Motivation for this thesis was based on Britest members requisite to gain fundamental understanding of the science behind cleaning. The main driver for this work is the potential benefits to Britest members. Therefore, Britest industrial members were actively involved in the provision of data and information present in this thesis. Industrial members contributing information included AMRI UK, Fujifilm Imaging Colorants Ltd, Hovione, Johnson Matthey, Pfizer, Robinson Brothers Ltd and Shasun.

1.5 Thesis Structure

Chapter 2, Considerations and Literature review, examines the industry requirements for process plant cleaning including aspects such as safety, environmental legislation, and regulatory requirements from bodies such as the FDA, which affect cleaning. This chapter also examines the contributing factors which underpin cleaning challenges and its complexity, including the effect of adhesion and cohesion of particles, solubility theories and mechanical aspects of soil or residue removal. Current cleaning understanding will be examined in the current literature available on cleaning and solubility theory. Industrial aspects of cleaning and the associated challenges will be considered with reference to a Britest member's survey, to begin to understand how to address the challenges and potentially identify solutions. Gap analysis of Britest tools and methodologies will be considered with a view to understanding, adaptation and developing for use in effective cleaning analysis and cleaning agent choice. In addition, cleaning metrics will be established to identify the benefits of further understanding of cleaning, based on the current industrial methods and solutions.

Chapter 3 identifies current science and theories which have helped to shape this research. This chapter discusses the industrial significance of this research by examining Britest members survey results relating to cleaning practises and challenges. In addition a number of site visits to Britest members are discussed which highlight their specific situations and dilemmas.

Chapter 4 begins to examine the knowledge and identify methodologies to solve the challenges associated with Britest member criteria. This requires analysis of theoretical and statistical methods to generate information and suggest solutions to the thesis questions. This chapter introduces and discusses attaining data and creation of databases for analysis.

Chapter 5 discusses analysis of the data by multivariate analysis, what using this method potentially indicates, and how it can be utilised in the design of a tool to aid choice of cleaning agent in industrial cleaning. Ultimately this chapter will begin to discuss design of a tool or suite of tools to be used with Britest members to facilitate a more scientific approach to process plant cleaning.

Chapter 6 demonstrates the use of adapted Britest tools for process cleaning and examines case studies where use has provided improved cleaning understanding. This chapter introduces the concept of FUSE, a suite of tools to assist and inform cleaning choices using a **F**undamental **U**nderstanding of **S**cience and **E**ngineering. Finally the chapter discusses and infers the implications and benefits for using FUSE.

Chapter 7 concludes the thesis by discussing the initial research question and summarises the conclusions. This chapter also presents further work.

Chapter 2. Literature Review

2.1 Introduction

This literature review aims to encapsulate the current understanding of industrial plant cleaning in a number of fields. Through this review, the application of fundamental science and engineering concepts behind cleaning are identified. This will help to determine accepted understanding and also identify gaps in current knowledge that can be investigated. Several areas were identified for the literature review, including Cleaning within the Dairy Industry (section 2.2.1), Cleaning and Removal of Food Particulates (section 2.2.2), Industrial chemical plant cleaning – ink and oil soil removal (section 2.2.3) and General cleaning information from industry and the biopharmaceutical industries (section 2.2.4). One of the fundamental questions this review aims to answer is - are there any commonly used cleaning methods or tools which can be applied to the pharmaceutical sector? It is considered that one of the most challenging questions for anyone considering cleaning in any sector is determining when clean is clean enough. It is therefore necessary to establish how this is determined and considered in this literature review.

Testing to ascertain whether equipment is clean is often carried out using analytical methods. This is the case in the pharmaceutical industry. The presence of residues and contaminants post cleaning are important considerations for this project, as they indicate inadequate or unsuccessful cleaning. Analytical methods are used for both equipment testing and also active pharmaceutical ingredient (API) testing for impurities, and can therefore determine whether cleaning has been successful. The current advances in analytical methods used to determine both validated cleaning and residual analysis will be considered in section 2.2.5. In addition to this, it is important to look at the regulatory stance with regard to cleaning in the pharmaceutical sector, and the implications this has on ensuring cleaning is carried out correctly. This review (sections 2.2.6 and 2.2.7) discusses several International Conference of Harmonisation (ICH) documents which are important in relation to pharmaceutical manufacture. The ICH is considered important as it ensures that standard medicines developed internationally are manufactured are safe, high quality and effective. The ICH also considers the associated levels of cleaning which are connected with the production of different types of active pharmaceutical product, or active pharmaceutical ingredients, as required by the regulatory bodies.

One of the major factors determining the effectiveness of cleaning, as determined by any analytical testing in any sector, is solubility. In section 2.2.8 the current understanding of solubility and its influence on cleaning, solutions and contaminants is discussed. Key solubility theories and models that help understanding of both process chemistry and cleaning challenges are indicated and their impact on this research is discussed. Similarly, the impact and relevance of identifying and defining chemical grouping, including by group contribution theory as a means to identify cleaning agents, is discussed in section 2.2.9. This has been considered by two means - classical chemical groups which are determined by chemical functional group, and secondly chemicals grouped by chemical reactivity. This will be discussed further in section 2.2.9.

2.2 Literature Review

2.2.1 Dairy Industry plant cleaning

The similarities between the bio-manufacturing and the dairy industry are apparent through the equipment required, the nature of the products made, and the need to control cleaning according to stringent regulatory requirements. The literature review will seek to determine the current cleaning problems and solutions in the dairy industry. By carrying out this review, challenges and solutions to current cleaning issues in the bio-pharmaceutical, chemical or pharmaceutical industries may be identified.

In the dairy industry, research has been carried out to determine the best methods to remove aggregated protein from industrial vessels and equipment, such as Bird, (1994), Bird, (1991), Bott, (1995), Burton, (1968) and Chen, (1998). The majority of the research carried out has concentrated on understanding why residues adhere to surfaces and how to remove them. It is acknowledged that cleaning is a major issue in the dairy industry. This is due to the frequency of cleaning that needs to take place. In the dairy industry emphasis is placed on making sure cleaning is effective to reduce the cost of repeated cleaning, and also to reduce the cost of the cleaning chemicals involved. Plant down time is a significant factor and it is important to reduce this and make sure the plant remains operational.

Research in the dairy industry on plant cleaning has focussed on a number of factors to help drive understanding and longer term improvements. One area includes the chemical composition of the residues, and the methods that influence their formation. Liu et al (2006a) described a technique to measure the adhesive strength of whey protein deposits. The work

concludes that the adhesive quality of the deposit is stronger than the cohesive strength. The contributing adhesive factors between surface and the deposit include van der Waals forces, electrostatic forces, hydrogen bonding and hydrophobic binding, together with contact area effects. It was stated that the greater the area, the greater the total attractive force. Liu et al (2006a) also describes the work of Visser, (1995) who stated that the adhesive forces can be reduced by the effect of surface hydration. Cohesive properties are thought to be due to covalent bonding. This paper indicates a good degree of understanding about the forces that affect soiling and soil removal.

Physical properties considered during soil removal have included both properties of the product and also how the plant is operated (Changani, 1997). The paper discusses the composition of the residue and the properties of the bulk fluid, in this case milk which has shown seasonal variation. Modelling for cleaning was carried out, with respect to chemical and physical properties. This was useful for comparing different cleaning chemicals. Grasshoff, (1999) determined that the rate at which cleaning was carried out was a first order rate constant but that this varied with time. It means that if a model is created for cleaning of deposits on surfaces it will be complex (Changani, 1997). This has implications for the development of a model for cleaning during this research, as it is likely that the creation of a model to determine the effectiveness of industrial cleaning will be complex and involve understanding of many factors. Other mathematical models have been developed to create understanding and predictability of cleaning. This includes work by Dürr (1999).

A mathematical model created by Dürr (1999) considered the removal of solid deposits on milk heat exchangers. This model indicated that factors such as flow mechanics, working time, composition, concentration and temperature of the cleaning agents, composition of the cleaning agents, composition of the deposits and the surface characteristics all influence the effectiveness of cleaning. Due to the factors involved which vary considerably, a mathematical model proved difficult to produce. However, a model was generated based on particle size of dust removal using a vacuum cleaner. This model was effective but it is not certain whether all of the above factors would have been included in this model.

Gillham, (1999), investigated the effects of cleaning in place (CIP), using alkali based solutions at a range of temperatures and flow rates. The results reported that cleaning comprised of three stages, these were a stage where the deposit swelled, a phase where the deposit began to uniformly erode and a stage where the deposit began to decay. It was shown that the rate of change in each stage was influenced by flow rate and temperature. The

research successfully showed the influence of the factors but it was not enough to enable modelling to take place.

Stainless steel cleanability in relation to the dairy industry has been studied by Leclercq-Perlat et al (1993). They investigated the effect of cleaning in relation to chemical modifications due to industrial cleaning procedures. The study reported that the cleanability of stainless steel types depended on a number of factors which included factors relating to the nature of the surface. This included the topography and the roughness of the steel. These factors influenced soiling due to the availability of attachment sites for soil or chemical bonding. The attachment of soil also depended on the properties of the soil. They indicated overall that soiling was less likely to occur on stainless steel that had a smooth finish on a microscopic scale. This is due to the fact that it has more hygienic properties, which it is likely to retain throughout its working life. They also thought that the use of a detergent that does not alter the surface will be the best for cleaning purposes.

Overall there has been a lot of research into fouling and soil removal in the dairy industry. This is useful in shaping ideas behind the mechanisms involved in chemical and pharmaceutical soil removal.

2.2.2 Cleaning and Removal of Food Particulates

With reference to industrial plant cleaning, it is thought that the largest body of relevant research lies within the removal of food soil other than dairy product, which has been discussed previously. The mechanisms of food particle deposit and removal are well discussed in literature. Many factors have been investigated which indicate why deposits occur and indicate methods and techniques for removal.

Durkee, (2006) described tasks involved in soil management as cleaning, rinsing, relocation of soil within the cleaning machine, drying and disposal of the waste soil. He states that in each of these tasks, '*soil is managed to produce a set of acceptable ends; part quality, productivity, disposal impact and operating costs*'. Durkee, (2006) explains that cleaning choices are based on the soils present or the equipment present. The depth of fundamental knowledge and understanding is not indicated in this work. The mechanism for removal is stated but there is no indication of why this is the best method, based on scientific or engineering principles.

Significant research has also been carried out to try and categorise, classify and define cleaning issues and types of fouling, as well as the mechanisms of deposit formation. The surface structure of the vessels used and the composition of the vessels have been evaluated with respect to the adhesive and cohesive properties of deposits, (Fryer, et al (2009)). These included both physical and kinetic properties (Simmons, 2007). Fryer et al (2009) identified a five stage mechanism which resulted in fouling. This mechanism briefly comprised of initiation, transport, attachment, removal and ageing. This was thought to occur for all soil types. The removal methods of soil are thought to vary according to soil complexity. A cleaning map indicting soil removal mechanisms shows the effect of cleaning fluid and the effect of mechanical removal. The effectiveness of both factors in conjunction is largely not considered. It is logical that both mechanisms can occur simultaneously and therefore should be considered. Methods researched in the field of food deposit removal make use of flow dynamics, temperature variations and the use of different cleaning fluids (Pritchard, (1988) and Van Asselt, (2002).

Liu (2006b) describes mechanisms of food particle removal. It is stated in the paper that food deposit removal is the result of failures in adhesive and cohesive properties. This means effective cleaning is influenced by these factors along with surface characteristics of the vessel and other unidentified factors. It was found that some food particles have stronger cohesive properties and others are more adhesive. This potentially means that different types of deposits could be removed effectively by targeted cleaning according to their adhesive or cohesive properties.

Liu (2006c) describes the identification and modelling of model food deposits by different modes. It is clear from this work that models can be produced for food deposit studies, showing how a food substance can fracture and break. The modelling carried out was simple and requires more development to understand what is happening at a fundamental scientific level, and also to determine what forces are causing the deposits removal.

It is clear from the preliminary literature review on removal of food deposits that there is a lot of unknown information relating to cleaning. The basic fundamental mechanisms of cohesion and adhesion are known but it is unknown why some deposits are more cohesive while others are more adhesive. The factors associated with soiling are many and this makes understanding what is happening difficult. Although modelling has been carried out on food deposit removal, it is not effective as yet.

2.2.3 Industrial chemical plant cleaning - Ink and oil soil removal

The ink and paint industries are good examples of industries that need to find fast effective solutions for the removal of soil. In the print industry removal of dried ink has a high cost impact due to the nature of removing soil from the microscopic cells on ink rollers. The cleaning method used needs to be effective, quick and leave the roller surface undamaged. Cleaning has generally been carried out by immersion in ultrasonic caustic baths, high pressure washing or manual cleaning which is ineffective and messy. However, successful methods to improve cleaning have determined supercritical mixtures of carbon dioxide and organic solvent (Della Porta, 2006). This operation indicates that properties of supercritical fluids are good at removing ink due to the density of the fluid and the viscosity, which falls in-between that of a gas and a liquid. The addition of an organic solvent to the supercritical fluid increases its cleaning ability. The only limitation of this is the fact that many solvents chosen for removal of soil have ignition temperatures close to those required for production of supercritical fluids under operation at the required high pressures. Della Porta (2006) indicates that supercritical fluids have a role in industrial ink removal as they remove ink quickly and effectively. Due to the properties of supercritical fluids indicated, high diffusivity and near zero surface tension may prove valuable in other fields as agents to remove difficult soil.

Supercritical fluids are of use in removing soil in the dry cleaning industry, as indicated by van Roosmalen (2004). During this application, mixing supercritical fluids with surfactants improved cleaning. When solvents were used in combination with supercritical fluids and surfactants, soil removal from textiles improved. This would indicate that the presence of the correct factors is the key to soil removal in any industry. Durkee, (2006), indicates that the cost of using high pressure equipment to achieve the conditions needed for cleaning in an industrial plant would be significant.

The removal of oil from vessels in industry is known to be challenging but due to the nature of the industry it is known that this is carried out infrequently. In the oil industry mechanical removal is the most common method of soil removal. This is often carried out by pigging in pipe work and by jetting and manual scrubbing in vessels, Harrington (2001). Manual cleaning is time consuming. It is important as with other industrial cleaning and maintenance programs that cleaning is scheduled into the manufacturing programme. Ishiyama (2010) describes the problems associated with fouling in oil refineries. A mechanism to prevent fouling was described and shown to be effective in a series of case studies. This involved the production of a simulation-based tool that was effective in controlling the desalter inlet temperature inside the boundary of a management strategy for foul control. The tool enabled

control of temperature alongside other factors to control fouling and it also emphasises the need for fouling control in the oil industries. It is thought that simulation and modelling similar to this could not be carried out in chemical and pharmaceutical industrial plants, due to the complex nature of industrial plant cleaning.

It is not known if chemical agents are used to remove oil from plant equipment on a large scale. Oil removal by chemical means has been shown by Al-Obeidani et al (2007). This was carried out on microfiltration membranes coated in oily seawater. It was found that a combination of alkaline and acid cleaning agents worked effectively, if the ratio of operating time and chemical cleaning time was increased and the amount of soaking time was reduced.

After a review of available literature it is apparent that the preliminary information found on ink and oil soil removal is of little relevance to this body of work.

2.2.4 General Cleaning Information from Industry and the Biopharmaceutical Industries

Cleaning is thought to be strongly influenced by factors such as flow rate and temperature. Work carried out by Cole (2010) indicates that removing type 1 soil (toothpaste) without using chemicals can be effective and that data collected during this exercise can be used to produce a model. Cole (2010) indicates that turbidity was used to monitor the removal of the toothpaste from a coupon in two locations while undergoing cleaning. The toothpaste was shown to leave the coupon in two stages. The two stages of removal were the core toothpaste removal and secondly a gradual removal of the remaining toothpaste. Modelling is currently underway to produce a cleaning model from this data.

It is thought by some that a good method for improving cleaning in the food industry is to use pulsed flow in pipe work. Augustin, (2010) states, '*A low flow oscillation imposed on a stationary flow of liquid has been shown to enhance sheer stresses imposed on a surface to mitigate fouling or enhance cleaning*'. Modelling indicates that this method may be of benefit in the food industry for removal of food deposits. It is not known whether this could work in the chemical or pharmaceutical sector. Prosek, (2005), also researched pipe cleaning using rinse based cleaning. In this study, pipe work of various geometries was designed, made and filled with a model solution. The work indicated that pipe work of different configurations was cleaned less efficiently in some cases. Pipe work that had valves or bends proved more difficult to clean. This study has implications on this plant cleaning project when considering the design of plant equipment in relation to the type of cleaning method used, the type of soil and many other factors. This is also a factor which concerns industrialists and will be discussed further in Chapter 3.

2.2.5 Analytical methods for determining residues and contaminants in vessels and in active pharmaceutical products

As the regulatory bodies require higher standards of drug purity and request more information on manufacturing processes, there is a greater need for improving analytical techniques in the pharmaceutical sector (Berridge, 1995). This is beneficial as it highlights purity, quality and gives an improved safety margin. It can also highlight residues, impurities in active pharmaceutical ingredients (API) and cleaning solutions left in vessels. It is necessary to identify, characterise and control any entity having Critical Quality Attributes (CQA) which has an impact on the quality of the drug substance (International Conference on Harmonisation ((ICH), Q11, 3.1.3). It is therefore important to consider the current and developing analytical methods which may be used to determine API purity, and also cleanliness of process equipment.

Currently the food and drug administration (FDA) are important in assuring public health as they ensure a level of drug, medical device and biological product safety, efficacy and security. The FDA also requires a level of cleanliness appropriate to the vessel used and drug type and potency. Drug impurity profiling is not a new technique but it is a critical task for pharmaceutical companies. Impurities above a detection limit of 0.1% are reportable in most circumstances (FDA, 1999). This means quantitation must be accurate. However, in some situations impurity levels of 0.01-0.1% are reportable (FDA, 1999). This level of reporting is applicable when there is the possibility of contamination with toxic or highly potent impurities. Currently several methods are used for impurity detection at these levels. These include chromatographic techniques such as high performance liquid chromatography (HPLC) carried out by many companies for isolation and characterisation of process related impurities. Several papers refer to this technique - Krishna Reddy (2002), Bharathi (2007) and Goverdhan (2009), who used the technique to detect impurities at levels below 0.10% in the drug Zafirlukast, used to treat pulmonary disorders. HPLC techniques have also been used to determine the effect of metal (copper and iron) degradation on the purity measurements of drugs by Dotterer (2011). It was found that the presence of less than 0.1 part per million (ppm) could lead to falsely low purity results. Spectroscopic methods such as mass spectrophotometry (MS) and nuclear magnetic resonance (NMR) spectrometry are used widely in industry as papers by Alsante (2001) and Roy (2002) indicate. Other techniques used are HPLC/ diode-array UV, Gas Chromatography/Mass Spectrometry (MS) and HPLC/MS (Görög, 1997).

Analytical techniques in use for characterising the quality of bulk pharmaceuticals have been questioned by some such as Görög (2005) who debates that the results of the final bulk drug assays give questionable results, leading him to believe that the drug quality measured is not what is achieved.

In carrying out this project confidential information on analytical techniques has been provided during the Britest member survey (Carr, 2011). It has been determined from this information that the analytical techniques carried out by Britest members are effective and they are also the only ones that are available to use in their circumstances. This will be discussed further in Chapter 3.

The area of analytical methods carried out for validation and verification of vessels and equipment used in processing post cleaning has been driven by industry standards. These include visual inspection, swabbing (which is then analysed by techniques used for drug impurity profiling, such as HPLC) and rinse water analysis. This was recognised during the examination of industrial survey results which will be discussed in chapter 3 (Carr, 2011). It is important to note that there are more novel methods of analysis under examination, some of these take into account surface properties of the materials of construction. This is an important consideration, as it has been determined by industrialists that material type, age and condition could affect the level of soil or cleaning residue accumulation in a vessel. In addition to this, the average velocity of cleaning detergents and the geometry of a vessel has been examined (Jensen, 2006). This describes the importance of the flow velocity of detergent as a factor in carrying heat and chemicals to clean vessels, and how changing this affects cleaning difficult plant geometry. Difficult plant geometry is described as crevices and dead-ends.

In the pharmaceutical industry standard practises drive the analytical assay types. This is important. As some companies strive to improve detection techniques, others must keep pace with new developments that, once established, become the techniques of choice by the regulatory bodies.

It is recognised in the pharmaceutical and biopharmaceutical industries that cleaning verification and validation take time to carry out. One reason for this is the length of time taken to analyse samples. In order to reduce this time, research is being carried out to provide solutions to reduce the time lag. Ion mobility spectrometry (IMS) has been used to determine residual API’s and their intermediates on equipment surfaces as a cleaning verification method utilising the normal swabs and rinse samples. This analytical technique is very

sensitive and able to determine quickly (in one minute) if a surface has residual contamination (Qin, 2010).

Another method currently in development by an American company, Block Engineering (2012), is designed to provide non contact real time verification of process vessel cleanliness. This is a technique that utilises infra red spectroscopy to determine the level of contamination from residual materials on the surface of stainless steel vessels. It is not known how the instrument deals with complex geometry found in vessels, but it has the potential to be a good tool for the assessment of cleaning. This would decrease the need for downtime while the vessel waits for analytical cleaning verification.

In addition to advances in analytical methods to help reduce cleaning time and increase effectiveness of detection, there have been advances in physical cleaning methods. This has included improving the method and operation of spray balls by Envirowise. This company states that it has advanced cleaning so effectively in the chemical, processing and manufacturing industries it has saved individual sites in excess of £80,000 year. (Envirowise, 2008)

Adhesion to different material surfaces including plastics such as polyvinyl chloride (PVC) floors and surfaces has been examined by several researchers such as Cramer (1972), Shebs (1987) and Pesonen-Leinonen (2006). Work has also been carried out on building materials, with modifications to enable the cleanability of easy to clean or self-cleaning characteristics to be determined (Määttä, 2011). This has shown the nature of these surfaces and revealed their nature using radiochemical methods, which are not generally applicable to vessels in the manufacturing industry, as the chemicals used are too hazardous.

Work carried out by Zayas (2006) and Resto (2007) has shown that it is possible to detect low level traces of detergents such as LpHse and CIP-100 in cleaned equipment. This was carried out by HPLC. In order to create industrial standards all assays developed and the results produced are subjected to regulatory inspection. This is critical in the pharmaceutical industry and of great relevance to this project, as can be determined from section 2.2.6.

2.2.6 Regulatory documentation and guidelines

In order to understand how the pharmaceutical industry is guided and monitored by the regulatory bodies, it is necessary to consider their documentation. Familiarisation with the regulations surrounding cleaning is an important aspect of understanding cleaning challenges. This is due to the fact that the development of a tool for use in the pharmaceutical industry

needs to incorporate an awareness of ICH and FDA regulatory guidelines. As there is not a lot of literature available on industrial cleaning, a valuable source of information on standard industry practises is found in regulatory guidelines. Regulatory guidelines for cleaning consider both physical and chemical aspects of cleaning. The acronym TACT; time, action, concentration and temperature is used to determine effective cleaning practises. It is thought that these factors play an important part in cleaning together and as a consequence they will be incorporated into cleaning tools developed in this research. The regulatory bodies consider that cleaning parameter selection is required when discussing cleaning and its effectiveness. Characterisation of biological entities in relation to cleaning challenges and different properties of soil is recognised. These factors must be understood in order to remove soil. The distinction between physical removal and chemical mechanism of removal is made by the regulatory bodies (Table 2-1).

Table 2-1

Physical Removal	Chemical Mechanisms
Static Soaking	Solubility
Convection	Emulsification
Dependant on soil size and degree of adhesion to surface	Wetting
	Chelation
	Dispersion
	Hydrolysis
	Oxidation

Table 2-1 Physical and Chemical Cleaning Factors (Adapted from Roessling, 2011)

Table 2-1 gives individual factors involved in cleaning but the factors can interact to clean.

If Fryer's (2009) work on cleaning contaminant characterisation and mapping is taken into account, then the cleaning challenges which are present in the biopharmaceutical industry lie within his map. Differences between the equipment and the chemicals used in the food, cosmetic and pharmaceutical industries mean that without further investigation it is not

possible to say in which zone entities belong. In order to help establish cleaning methods and improve the understanding of cleaning factors which interact, experimenters use coupons.

Cleaning experimental design can be discussed with reference to the use of coupons of known materials and surface types. The coupons can then be impregnated with a know amount of soil and cleaned to determine the effectiveness of the cleaning and the percentage soil recovery. This is used in industry to determine cleaning effectiveness. However it is difficult to get a truly representative coupon with regard to age and surface roughness. Coupon use must always take this into account. Application of design space for cleaning experiments and the importance of identifying risks within this space is key. Cleaning limits with regard to vessels and also limits of contaminants and carryover into bulk drugs should be considered. Methods used to determine the maximum allowable carryover limit are discussed with reference to ICH guidelines. Firstly, it is important to discuss the need for identification of potential risk factors and their control (deemed critical process parameters (CPP)) and the impact to cleaning processes. Critical quality attributes (CQA) are factors used to determine the effectiveness of the cleaning activity. It is thought that a list of both CPP and CQA can be created for incorporation into a cleaning tool which is generic to the pharmaceutical industry, but specific enough to fulfil company requirements. Another consideration which must be addressed when designing cleaning tools is the influence of the regulatory bodies.

2.2.7 International Conference on Harmonisation (ICH) regulations

In industry, pharmaceutical processes are generally well understood by companies who use them. Process characterisation has the benefit of ensuring products are produced to the correct specification, the right quality and the best yield. However, full characterisation of processes is not generally carried out. This is due to several factors which include time and resources. It is not thought to be beneficial to fully characterise a process in industry and impurities are only considered if they are present in large amounts and limit the process yield. They may also be characterised if they are thought to interfere with the safety or the efficacy of the active pharmaceutical ingredient (API).

If impurities are not fully characterised in enough detail there is a risk that products can carry unknown impurities (unwanted chemicals that stay with an API or develop with it during processing). This is a critical issue and, in order to address and control this risk, the International Conference on Harmonisation (ICH) has produced several guidelines in order to control impurities. Many of the ICH guidelines have influence over this body of work and these are listed in table 2-2 which describes the relevance of each document to this project.

Table 2-2

ICH Document Reference	Relevance to Industrial Cleaning Project
Validation of analytical procedures: Text and Methodology Q2(R1)	Guideline relates to the validation of methods for use in identifying and quantifying contaminants, impurities and residues. Advises on assay validation characteristics such as accuracy, precision, repeatability, intermediate precision, specificity, detection limit, quantitation limit, linearity and range
Impurities in new drug substances Q3A(R2)	Impurity classification (solvents, organics and inorganics) and reporting.
Evaluation and recommendation of Pharmacopoeial texts for use in the ICH regions on Test for particulate contamination: Sub-visible particles general chapter Q4B ANNEX 3(R1)	Guidelines on testing for particulate contamination: Sub-visible particles
Specifications: Test procedures and acceptance criteria for new drug substances and new drug products: Chemical substances Q6A	Guidance on setting and justification of acceptance criteria and selection of test procedures for new drug substances of synthetic chemical origins and new drug products produced from them.
Good manufacturing practise for active pharmaceutical ingredients Q7	Guidelines on operating within Good Manufacturing Practises (GMP). Covering personal, quality management, buildings, process equipment, documentation, facilities management, storage and distribution and production and in-process controls.
Pharmaceutical Development Q8(R2)	Guidance on pharmaceutical development, designing a quality product and manufacturing process.
Pharmaceutical Quality Systems Q10	Guidelines for establishing a quality system
Development and manufacture of drug substances (Chemical entities and Biotechnological/ Biological entities) Q11	Guidelines for developing and understanding manufacturing process of a drug substance.

Table 2-2 ICH Documentation Relevant to the Industrial Cleaning Project (Adapted from ICH documentation). (ICH Q2 to ICH Q11 guidelines)

There are two documents listed above (ICH Q3 and ICH Q11) which bear more relevance on this research than the others. It is necessary to briefly explain their relevance.

ICH Q3 as listed in table 2-2 discusses impurities in drug substances. Analysis of API products can show contamination. It is important to understand where this comes from to minimise it and potentially prevent it from happening in future batches.

One of the most important areas which can lead to contamination of drug products is a lack of process understanding. This can take many forms, such as a lack of understanding of the process chemistry giving rise to contaminants, for example from side processes. Impurities may also enter products from solvents used in the processes, including cleaning solvents, if the correct cleaning techniques are not adopted. It is possible that impurities may be left in vessels from one manufacturing process which can then contaminate the following process. Issues like these highlight the need for greater scientific understanding, including during cleaning processes. ICH Q3 details guidelines for validating analytical procedures to help determine impurities, and also gives limits for the detected impurities. Impurity limits are determined by the type of impurity, the toxicity and the amount present. The document gives guidelines on this.

ICH Q11 is a guideline which details how understanding the manufacturing process can lead to consideration of impurity development and how this can be minimised. The guideline describes two methods for developing a drug product. The first is the traditional method and the second is described as an enhanced method. The traditional method determines the development of a product within a series of set points and operating ranges for process parameters. The strategy is to control the process and show it is reproducible and can meet determined acceptance criteria. This is commonly used in industry.

The enhanced method is more detailed and examines risk management and scientific knowledge to provide an in-depth understanding of process parameters and operations that can influence critical quality attributes (CQAs). This provides information for the development of control strategies which can be used over the entire process. This can include creating a design space. A design space is identified as the ‘multidimensional combination and interaction of input variables (e.g., material attributes) and process parameters that have been demonstrated to provide assurance of quality’ (ICH Q11, section 3.1.6). If operations are carried out within the design space the process is not considered as changed and does not require regulatory post approval change process.

Guidance is given on linking material attributes and process parameters to drug substance CQAs. This is relevant to the research carried out in this project as it advises the development of a design space based on using prior knowledge (experience of the chemistry and the

process) and also chemistry first principles. This indicates that process understanding should be based on scientific understanding. Understanding the chemistry behind cleaning is therefore critical. In order to carry this out the influence of solubility theories are examined in section 2.2.8.

2.2.8 Solubility Theories and Models

As determined in section 2.2.5 there are a number of factors which can influence and affect cleaning. In order to understand cleaning it is therefore important to consider the impact that solubility has on cleaning solutions and contaminants. This has also incorporated current Britest members understanding of solubility in relation to their cleaning operations, which will be discussed in chapter 3. Solubility of chemicals during processing, analysis of product samples and chemical residues (soil) and in aqueous agents and solvents has not been extensively studied by researchers. It has however, been determined that solubility is not only a critical aspect of understanding how to improve processes and product manufacture, it also vastly improves the ability to clean equipment post processing. In order to understand solubility it is necessary to establish the importance of choosing parameters. These parameters have been discussed by Durkee (2004a), where he states that there are ways to determine how to match the best solvent to a soil. Durkee maintains that by understanding systems such as the Kauri Butanol (Kb) test, the Hildebrand Solubility Parameter and the Hansen Solubility Parameters, this can be achieved.

Kauri Butanol was the primary test used to typify the dissolving power of a solvent. It is a measure of solvent strength. This indirect test involves using thick Kauri gum (a hardening agent in varnishes and enamels) as soil and determining if a solvent can solubilise it. This produces a solvent Kb value. The test is only useful if the soil has similar properties to Kauri gum. This test can only be used for hydrocarbon solvents such as Benzene or Toluene. It is not used for polar solvents such as esters or alcohols and the values produced for these solvents have no meaning (Baldeschwieler et al, 1935).

The Hildebrand Solubility test is a method of characterising solubility in solvents (Hildebrand et al, 1949). It was proposed by Hildebrand that energy is required in order to overwhelm intermolecular forces. Simplified this means that ‘like dissolves like’ (Durkee, 2004b). The best way of calculating how much energy is required to solubilise soil is to determine how much energy is required to vaporise the solvent (Durkee, 2004b). One dimensional solubility is based on all intermolecular forces within a solvent, such as Van der Waals forces, polar interactions, dispersion forces and hydrogen bonding forces.

Hansen expanded on the ideas of Hildebrand, among others, and divided the Hildebrand solubility parameter into a three dimensional system. Each part was associated with an intermolecular force (Durkee, 2004b). This created advantages in understanding and predicting solubility as some molecules within solvents are driven by some forces with more strength than others. Hansen's solubility factors are one of the most established and accepted solubility theories, but an important consideration when using Hansen's parameters is that they must be known for a soil as well as the solvent (Hansen, 2007). Work is still ongoing to determine and enhance knowledge in this area, using techniques such as quantitative structure activity relationship modelling (QSPR), combined with Conductor like screening model (COSMO). This has provided information that can be used to help characterise solvents at a molecular level and further understanding of solubilisation (Járvás, 2011).

There are many other systems for understanding parameters around solubility and the choice of a solvent in relation to materials, such as the work by Aharoni (1992) on swelling measurements, Adamski (2008) using inverse gas chromatography, Roberts (1993) using mechanical measurements, Bustamante (2000) utilising solubility/miscibility measurements in liquid with known cohesive energy and Gharagheizi (2006) by viscosity measurements.

Solubility is not the only aspect of chemistry influencing this research. Important and relevant fundamental chemical interactions also need to be addressed. It is important to consider the relevance of identifying and defining chemical grouping. In order to achieve this, chemical functional groups and reactive groups are examined in section 2.2.9. This has been considered by two means - classical chemical groups determined by functional group, and also by chemical reactivity.

2.2.9 Chemical functional groups and reactive groups

Significant work has been undertaken in order to determine a system to group chemicals in organic chemistry, in order to predict behaviour and help understand their properties. Chemicals can display huge variations in their properties and the way they act and react together. Two major systems of classification have been developed. The first is to group chemicals according to reactivity (devised in the 1930's), and the second is to group chemicals based on their chemical functional groups. Chemical reactive grouping is based on the fact that typically chemicals react in similar ways due to the fact that they have similar chemical structures. These groups are identified in table 2-3.

Table 2-3

Reaction Type	Example
Nucleophilic Substitution	Alkyl Halides
Electrophilic Substitution	Alkyl Metals
Radical Substitution	Alkanes
Nucleophilic Addition	Aldehydes, Ketones and Nitriles
Radical Addition	Alkenes, Alkynes
Nucleophilic Addition - Elimination	Carboxylic Acid Derivatives
Electrophilic Addition - Elimination	Arenes
Elimination	Alkyl Halides
Pericyclic	1, 3 Dienes

Table 2-3 Reaction Types and examples. (Adapted from information supplied in Massey, 1990).

As the information in table 2-3 suggests, there are a number of reactive groups, and several chemical types are found in more than one group, as they have more than one reaction type. The grouping is easy to understand and there are only fifty groups, which is an advantage over the second method, grouping by chemical group.

The second method, or chemical grouping, is by functional groups and was the first method devised in order to group chemicals. Functional groups are based on organic chemistry (Inorganic chemistry has the same functional groups but, it is the elements which dominate the chemistry and the functional groups perform a moderating function). There are known to be over one hundred groups identified by this method. Some of these groups are listed in table 2-4.

Table 2-4

Functional Group	Definition
Alkenes Alkynes Aromatics	Contain C + H only
Nitriles Amines Amides	All contain Nitrogen
Alcohols	Involve a single bond which contains Oxygen
Phenols	Involves double bonds which contain Oxygen
Ethers	Involve a single bond which contains Oxygen
Aldehydes Ketones Carboxylic Acids Acid Chlorides Acid Anhydrides Esters	Involve double bonds which contain Oxygen
Thiols Thiol ethers	Contain Sulphur

Table 2-4 Chemical Functional Groups Information. (Adapted from Housecroft, 2005).

It is this method of grouping chemicals which will be utilised for this research. Chemical grouping, identifying key characteristics, was chosen in order to group common chemistries for this project. This is because the method is easy to understand and the grouping structures are easy to explain to industrialists. There is also a lot of literature available on this methodology. The common chemistries including those described in table 2-4 will be used for this research project moving forward. This information and how it will be used in this research is further discussed in Chapter 4. It is also important to consider group contribution methods which may bear relevance to this work. This will be discussed in section 2.3.

2.3 Group contribution methods

Methods for predicting the behaviour of groups of chemicals are known as group contribution methods. There are many forms of these methods which all have slightly different reasons for groupings. It is well known that several of these methods are known to be inaccurate and

unreliable (Constantinou, 1994) and therefore they may not be very useful for this project work. Methods of group contribution include grouping by primary properties (for example molecular structure, critical pressure, normal boiling point and Gibbs energy) and by secondary properties. Another class of group contribution methods are based on the type of functional groups such as Benson Group Increment Theory (BGIT) (Benson, 1958). This is known to be a complicated method of group contribution method (Constantinou, 1994). Another group contribution method is universal functional activity coefficient model (UNIFAC). This is a method which uses functional groups present on the molecules that make up liquid mixtures to calculate activity coefficients. It was devised to predict interactions between molecules ,by describing occurring molecular interactions based on the functional groups present on the molecule, (Fredenslund,1975). In the pharmaceutical industries this is a recognised and commonly used method for predicting interactions and activities of molecules.

Methods by analytical solutions of groups (ASOG) are also used, for example in order to determine water activity in solutions of sugars and urea (Correa, 2004). Some researchers indicate that UNIFAC and ASOG methods can have weaknesses (Gmehling, 1993). Group contribution methods have been used for predicting the solubility of solutions by several researchers for seed polyphenols of vitis vinifera (Savova, 2007) and non-electrolyte organic compounds (Gharagheizi, 2011). Water solubility of organic chemicals was estimated by group contribution methods (Kühne et al, 1995) and the solubility of selected pharmaceuticals in both aqueous and non aqueous solutions was predicted with a group contribution method (Pelczarska et al, 2013) with a degree of success. Therefore it is possible that the use of a similar method for determining groups for industrial cleaning chemicals, residues and contaminants may be used. This is because solubility of pharmaceuticals in solvents used for cleaning is a similar challenge to solubilising pharmaceuticals for drug manufacturing and delivery purposes.

2.4 Chapter 2 Summary

In Chapter 2 it has been possible to examine some of the current literature and theories which will help to shape and guide this research. The literature review has shown that the most commonly published information in the area of cleaning is the dairy and food industries, where there are some key similarities. These similarities include difficult to remove entities, complex plant or equipment and often fouling by similar residues. The literature review also determines that cleaning is often specific to a process or a type of equipment. There are

differences in how cleaning is approached in sectors using combinations of key removal factors such as physical removal, chemical removal or a combination of both.

In addition, the literature review indicates that the use of group contribution methods as a method to help group cleaning agents may be effective, but it should be approached carefully due to flaws in many methods. The most apt method for grouping pharmaceuticals and API in this research may be the UNIFAC method. The advantage of this is that it is a widely used method in industry and any cleaning models created during this research will be easier for industrialists to understand and implement.

The literature review on the subject of industrial cleaning itself helps to determine one factor moving forward in this research. That is, there is not a lot of literature in the public domain on pharmaceutical plant cleaning. It can be concluded that the best, and significantly the most important way of examining current cleaning practises, is by direct contact with industrialists. This is discussed in Chapter 3, where approaches to cleaning are discussed with the help of a survey aimed at industrialists on a variety of plant. Chapter 3 will also further examine the economic reasoning behind this research using defined metrics gained from industrial understanding of pharmaceutical plant cleaning. As the aim of this project is to determine or design a methodology or tool for pharmaceutical plant cleaning which fits into the Britest remit, it is also important to consider the remit of Britest tools in Chapter 3.

Chapter 3. Industrial Considerations

3.1 Introduction

This chapter identifies current science in industry and theories which have helped shape and position this research in addition to the academic literature review in Chapter 2. It is important to consider the industrial significance of this project and establish why it is important to conduct this research. This was carried out with the aid of a cleaning survey and site visits to Britest members. The survey helped gain an overview of the current challenges and approaches taken to solve them (section 3.2). Site visits gave valuable insights into the engineering and cleaning challenges that Britest members are facing (section 3.3). This chapter will help position the research by answering RQ3, RQ4 and RQ5 (section 3.4). It is known that the nature of cleaning is complex and factors which determine effectiveness can vary due to the nature of the equipment and the products made in the equipment. At this stage in the research it is critical to introduce and review the Britest tool set (section 3.5). This will help to determine if any existing tools and methodologies can be used in their current format, to help Britest members understand cleaning challenges and how to improve plant cleaning methodologies.

Furthermore, the cost of cleaning will be assessed in this chapter, in sections 3.6 and 3.7, to establish how implementation of potential findings of this research could save industry time, financial cost and resources. Finally, section 3.8 will summarise this chapter and indicate the direction of the research in Chapter 4.

3.2 Industry Requirements contributing to the Research

This section aims to answer the following research questions by engaging with Britest members by survey and site visits to view plant equipment:

RQ3: What are the main challenges associated with process plant cleaning for Britest members?

RQ4: What common methods do Britest members utilise to clean their process equipment?

RQ5: What cleaning challenges are associated with plant engineering or choice of cleaning agent?

As stated in Chapter 1, section 1.4, the need for this research is underpinned by Britest member's requirement for a fundamental understanding of the science behind cleaning. In order to understand exactly what is meant by this term it was necessary to ask Britest members some questions. This was carried out by engaging Britest members in a survey on process plant cleaning.

Prior to considering the results of the survey, it is important to determine what is meant by the term 'plant', as used in the context of this research. The term plant indicates a general term to describe the processing or manufacturing equipment used by Britest members in the pharmaceutical, fine chemical or chemical industry. It is necessary to state that this equipment is generally complex. The complexity of equipment derives from the ages and types of equipment used (which are described below). The number of processes which are involved to manufacture a product make it quite complex and often the system of pipework used to connect the equipment is not designed specifically for purpose, leading to deadlegs in pipes which are difficult to clean. In addition the layout or design of the equipment may be determined by current use, or by historical use. Plant tends to be different dependent on the product manufactured. In many companies the majority of plant is composed of similar classic unit operations and equipment. For the purposes of this research typical plant equipment and unit operations are identified as reactors and holding vessels, separation equipment such as centrifuges and chromatography columns, filtration equipment, drum dryers, vacuum dryers or spray dryers. In addition, raw material handling equipment, such as glove boxes and powder handling systems are considered typical plant. Britest member plant also includes pumps, associated pipe work, and commonly, condensers and heat exchangers. As plant can be configured differently at different companies, it is important to understand individual Britest members cleaning challenges. In order to achieve understanding, Britest members were surveyed.

An initial survey carried out (Talford, 2009) indicated there was a need for increased cleaning understanding and indeed cleaning awareness. Britest industrial members felt that plant cleaning protocols were developed in an ad hoc fashion, with little consideration or understanding of the science behind what they were trying to achieve. It was indicated that cleaning methods were developed on the basis of trial and error. This mode of operation gave an intolerable level of cleaning failures and consequently it resulted in a number of negative impacts. These include cost of cleaning agent use and disposal, missed scheduling times for batches and ultimately a financial cost in plant down time or lost opportunity. The primary

target for plant cleaning is to achieve right first time (RFT) cleaning. Britest members indicated that there was a demand for increased fundamental scientific knowledge around cleaning and specifically wanted Britest tools to help understand and solve cleaning challenges.

This information is considered important, but in order to gain more insight a second survey was commissioned for this research project (Appendix 1 - Survey questions). There were 11 responses to this survey from different pharmaceutical and chemical companies (There were 18 member companies at this time).

The second Britest member's survey on cleaning was specifically aimed to answer RQ3, RQ4 and RQ5. The aim was to understand process cleaning with respect to the following -

How is a plant cleaning protocol developed?

Are there any specific plant cleaning issues?

What are the current plant cleaning methods?

Which cleaning agents are used on site for cleaning?

What effect do analysis time/cost and validation requirements have on cleaning?

What are the financial and time implications of ineffective cleaning?

The survey key findings indicated Britest members developed plant cleaning protocols largely based on contaminants present (Figure 3-1). Given that there is a need to ensure that cleaning is carried out efficiently but rapidly, this is understandable. The question does not however, indicate the depth of contaminant understanding.

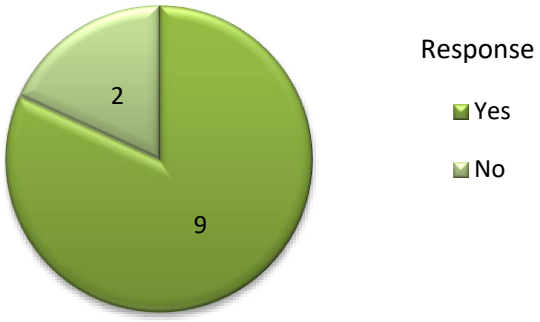


Figure 3-1 Are Cleaning Protocols based on understanding contaminants?

Further to this Britest members understood the types of contaminant or residues present in the vessels and equipment. Chemical contaminants or residues are indicated as the main objectives of removal (Figure 3-2).

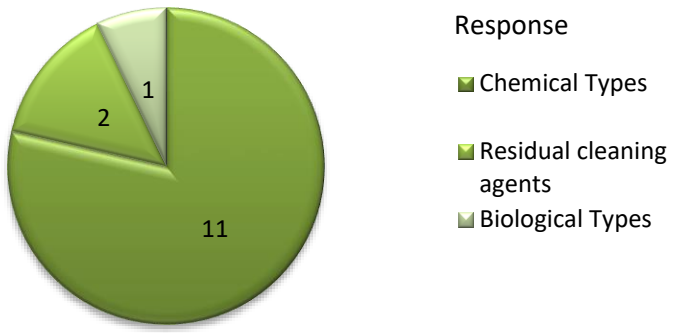


Figure 3-2 What is the main contaminant type in your process?

With the main contaminants known, cleaning should become easier as cleaning agent selection can be targeted. In spite of this, there are several other factors to consider when devising a cleaning protocol. These factors include solubility of the contaminant and mechanical action (Chapter 2). Figure 3-3 indicates that there is generally more than one factor influencing cleaning protocol decisions for Britest members.

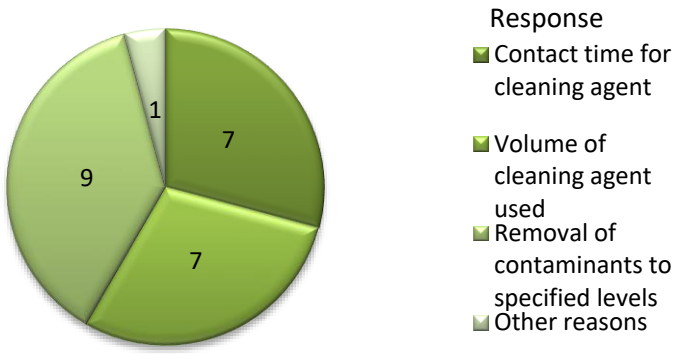


Figure 3-3 Factors influencing cleaning protocol design.

Figure 3-3 indicates the level of complexity around cleaning and shows a need to identify all factors involved in the design of a cleaning protocol if it is to achieve RFT cleaning. In addition to Figures 3-1 to 3-3, a number of qualitative answers to questions were received,

indicating how Britest members create cleaning protocols. This suggested that a majority of companies surveyed used a very simple method to select cleaning agents. This will be discussed later in this section. Qualitative answers were informative as they set the scene for plant visits, gave questions for targeted interviews and indicated potential cleaning case studies (section 3.3). They also highlighted common cleaning challenges associated with equipment, such as those surrounding condensers. Cleaning methods varied between Britest member companies. This highlighted the fact that a lot of cleaning methods are unique to companies and often specific sites within those companies if there are multiple plants. The reason for this is not known. It could be considered that a well run optimised plant cleaning process gives a competitive advantage over others if it works RFT, therefore cleaning information is not commonly available in literature, or willingly shared. It is thought that by improving plant cleaning the cost of drug products could be reduced, due to decreased overheads associated with cleaning. As the survey was aimed at all Britest members it is important to state that the member companies can be categorised according to different factors, such as size, pharmaceutical production and fine chemical production, and therefore the type of cleaning required (Table 3-1).

Company	**Multiple site**	**Type of manufacture**	**Type of cleaning required*[1]**		
			Validation	**Verification**	**Other*[2]**
Johnson Matthey	Y	Chemical/ Pharmaceutical	Y	Y	
Pfizer	Y	Pharmaceutical	Y	Y	
Fujifilm Imaging Colourant	Y	Chemical		Y	Y
Hovione	Y	Pharmaceutical	Y	Y	
Albany Molecular Research	Y	Chemical/ Pharmaceutical	Y	Y	
Robinson Brothers	Y	Chemical		Y	
Isochem	Y	Chemical	Y	Y	
Chemie Uetikon	Y	Chemical/ Pharmaceutical	Y	Y	
Astra Zeneca	Y	Pharmaceutical	Y		

Table 3.1 Britest member participation in the cleaning survey.

*[1] Indicates type of cleaning required which depends on type of product manufactured.

*[2] Indicates it may be possible to determine clean equipment by stating visually clean.

AstraZeneca submitted more than one set of survey results based on the findings of two different departments.

In addition to this, cleaning can also depend on the type of product manufactured and the type of product, which is made after the vessel is cleaned. For example, if a batch of product is made in multipurpose equipment which has been used for a toxic or potent product, then it may be expected that the cleaning carried out on the equipment is quite extensive.

The cleaning survey implied that process plants are often made up of the same equipment, but the equipment can differ significantly in age, operational use and design. This makes plants dissimilar. The cleaning equipment used at each plant varies significantly and it is often common to target specific areas for special cleaning. These areas can vary, between batches of different product and batches of the same product in the same vessel. Key survey findings revealed that contamination does not occur in one specific part of process plant. Methods to target specific areas for cleaning can thus vary (Figure 3-4).

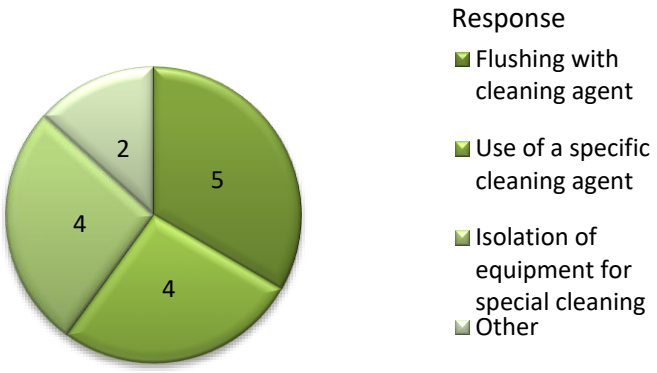


Figure 3-4 How is an area targeted for specific cleaning?

Figure 3-4 gives a number of responses to the question and shows that companies will adopt more than one approach to attempt to get equipment clean. This indicates a trial and error approach. Companies consider it appropriate to bring in specialised equipment or cleaning companies for process cleaning challenges. Therefore, placing the cleaning responsibility in someone else's hands is easier than cleaning themselves. This could render equipment out of use or cause a schedule hold up if cleaning specifications are not met quickly. Figure 3-5 indicates that the cleaning methodology used is based on solubility of the contaminants, as

specific cleaning agents are used to solubilise a contaminant rather than choosing an agent based on known molecular structural properties. The selection of a cleaning agent is currently carried out by trial and error in the majority of companies surveyed. This can be achieved by using a simple solvent screen carried out for the reaction chemistry. It does not however, take into account the solubility of potential contaminents.

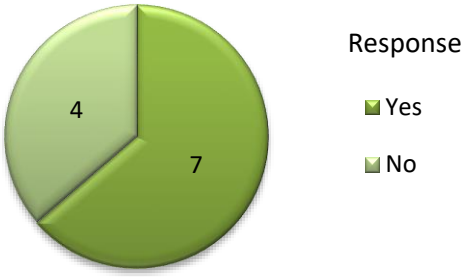


Figure 3-5 Have you identified any biological or chemical structure which can be targeted by the inclusion of a specific cleaning agent?

Where yes was indicated (Figure 3-5) it is necessary to state that the structure or contaminant is targeted by an acid or alkali cleaning agent. The methodology used here points towards trial and error with companies selecting cleaning agents until they achieve the most effective cleaning, which is not optimised. The effectiveness of cleaning protocols was assessed during the survey. The results showed that a majority of the time cleaning protocols were considered sometimes effective (Figure 3-6).

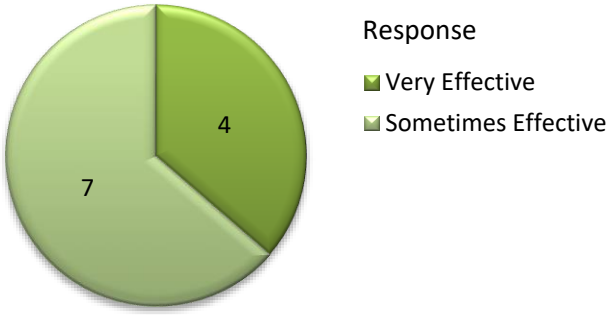


Figure 3-6 How effective is your cleaning protocol?

The responses showed no protocols were completely effective or ineffective. It is not known how Britest members define effective. If ineffective cleaning is found, cleaning is repeated until it is satisfactory. This may mean cleaning using the whole protocol again in full, repeating parts of the cleaning, or isolating process equipment as necessary for special cleaning. This may mean that cleaning documents are not as defined as other process documents. It is certainly true for verified cleaning protocols. Britest members implied that a range of cleaning agents are used (Figure 3-7). This indicates that the agent varies with the process or product manufactured. The use of cleaning with detergents other than aqueous detergents was not indicated in the survey.

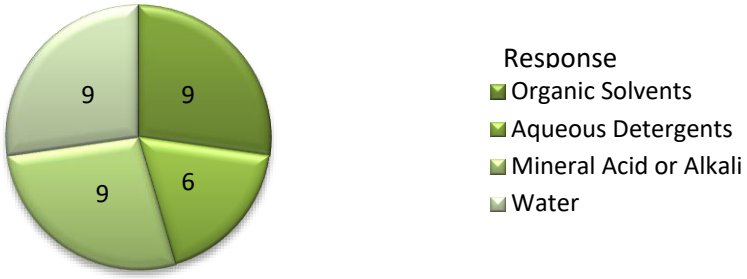


Figure 3-7 Cleaning agents used by Britest members.

Most respondents indicated they used water to clean equipment in addition to other agents such as organic solvents, acids and alkali. Respondents were asked to indicate how they selected their cleaning agents (Figure 3-8).

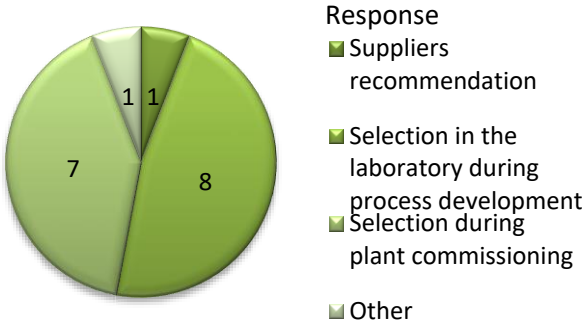


Figure 3-8 Criteria used by Britest members to select cleaning agents.

Figure 3-8 indicates that the selection of cleaning agent is not an early consideration during the development process. Often the decision on which cleaning agent to use is made during the late product development. This can be problematic. If cleaning were to be considered as part of whole process design, it is possible that this may influence the selection of the process chemistry. A less elegant process chemistry may give sustainability benefits once cleaning was factored into the life cycle assessment. Solvents for cleaning purposes are chosen based on solubility studies and not fundamental science. The method of choice in industry is best shown figuratively (Figure 3-9).

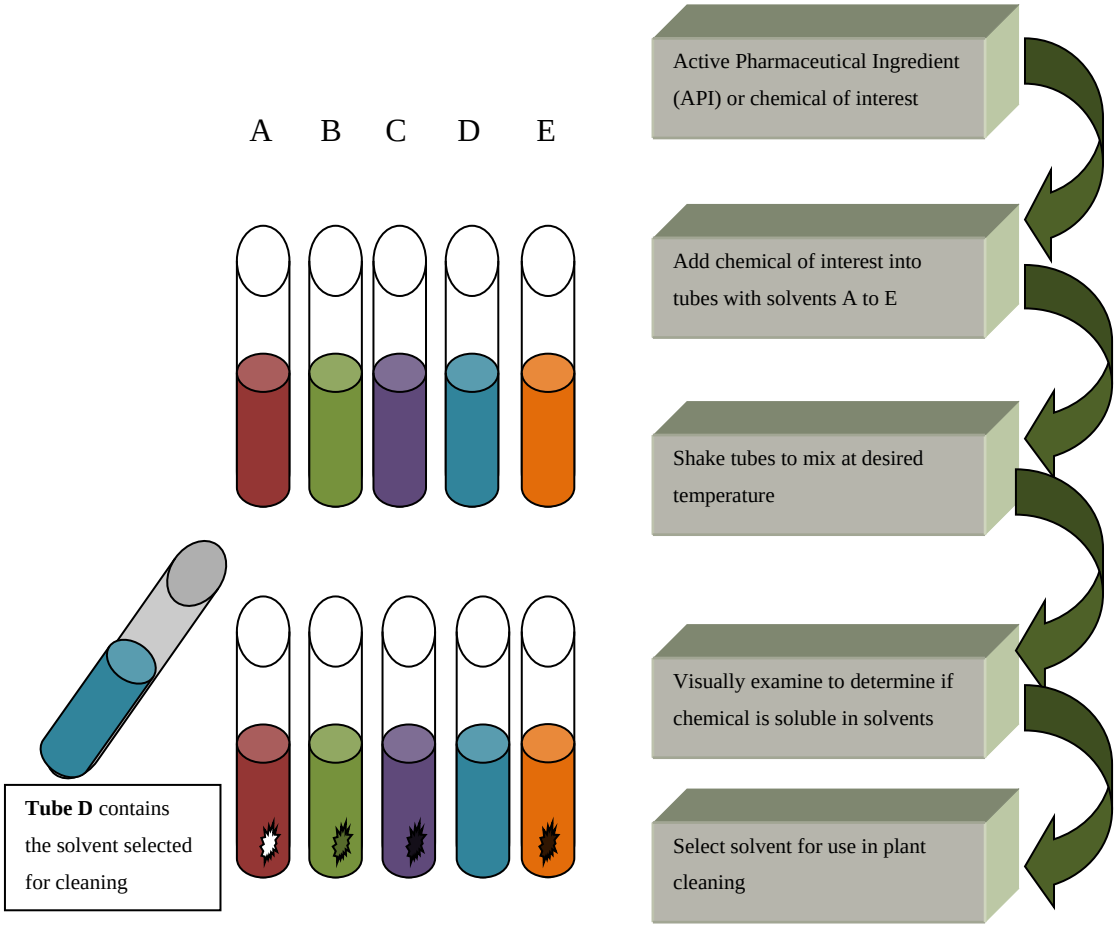


Figure 3-9 Industrial method for selecting a cleaning agent as provided by Britest members. Where tube D is the best choice of solvent with no visible residue remaining.

Figure 3-9 shows the method of solvent selection in industry. This is not considered the most rigorous or scientific based methodology, but it gives an indication of the best solvent to use. It must be remembered that industrial methods such as the one described above only considers

the key reagents, not all the potential products as contaminants. The impact of solubility on determining the direction of this research is considered in Chapter 2 (section 2.2.8).

Britest members indicated that this method is not an adequate method of choosing a solvent for cleaning purposes. This is also apparent in the response to questions regarding effectiveness of cleaning (Figure 3-10). Respondents had varying degrees of success with their cleaning protocols.

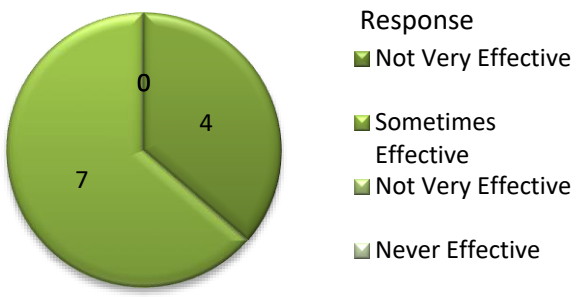


Figure 3-10 Effectiveness of current cleaning protocol?

Figure 3-10 indicates that although the effectiveness of cleaning is tolerated by the industrialists who were surveyed, right first time cleaning was rarely achieved. This survey question highlights evidence which indicates that the cleaning protocols in place are not optimised. Good Manufacturing Practice (GMP) guidance is given on cleaning post manufacture and pre-manufacture, which determines the recommended parts per million (ppm) contaminants in drug products, drug product intermediates and in vessels used for manufacturing dependant on quality control and quality assurance departments with organisations (as referred to in Chapter 2 sections 2.2.6 and 2.2.7). In order to achieve this, without significant lost time and waste cleaning agent due to repeated activities, cleaning must be optimised. It is considered that the main impact of this situation is on process scheduling and the financial implications of this. If the effectiveness of cleaning is tolerated, the next questions which must be asked is - what are the range of cleaning agents in use by Britest members and how are these selected? Figure 3-11 and 3-12 indicate the answers to these questions.

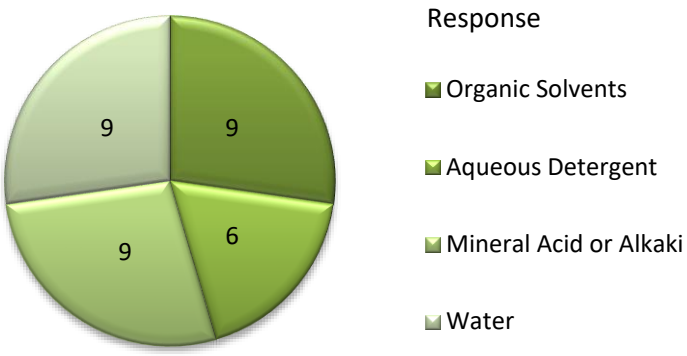


Figure 3-11 What cleaning agents do you use?

Figure 3-11 implies that the types of cleaning agents used by Britest members are both aqueous and solvent. Three companies indicated that they did not use aqueous detergent. If companies are using more than one type of cleaning agent, how are they selecting those cleaning agents? Figure 3-12 indicates the answer to this question.

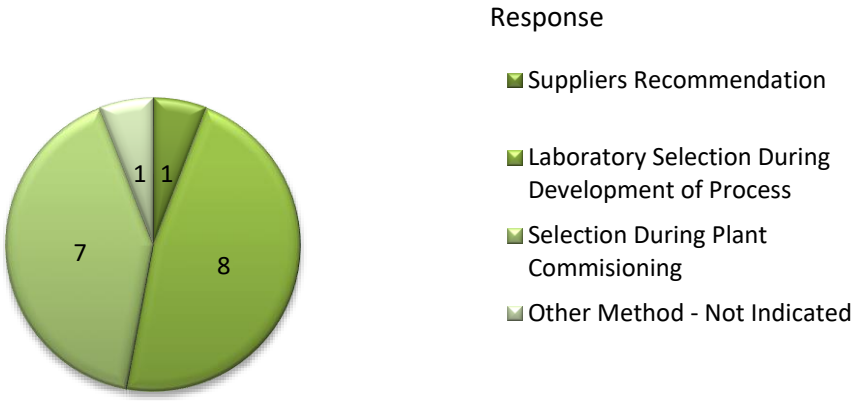


Figure 3-12 How are your cleaning agents selected?

Figure 3-12 shows that the selection of solvents is mainly determined by two factors which are firstly a laboratory selection during process development, as indicated by figure 3-9. This method of choosing a solvent is not adequate, as the survey data related to cleaning effectiveness indicates (Figure 3-10). Right first time cleaning is not often achieved. In addition, the methodology used to determine solubility does not take into account the different surface types, ages and equipment which needs to be cleaned in processing plant. The complexity of this equipment is difficult to simulate in a laboratory environment. Solubility of

the worst case product is used in the selection of a cleaning solvent. This does not consider the importance of intermediate and side products which may form in processing and prove more difficult to remove than the product used for the solubility test. The second most favourable choice for solvent selection is during plant commissioning. At this stage in the life-cycle of process design it is very late to change a manufacturing process. This may mean that a process is introduced into a plant with ineffective cleaning processes designed too late in the process life-cycle to change. The survey results overall indicate that cleaning considerations occur later than should be expected if whole process understanding is to be applied and this lacks fundamental scientific understanding of the cleaning processes employed by the Britest members. Figure 3-13 shows in more detail the effectiveness of Britest members cleaning protocols with regards to cleaning carried out right first time.

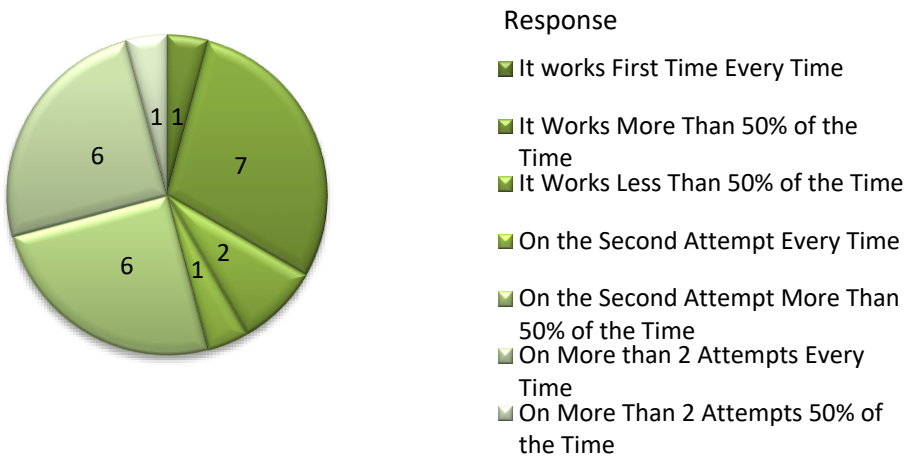


Figure 3-13 Can you state the Effectiveness of your current Cleaning Protocol?

Figure 3-13 suggests that the cleaning protocols work 50% of the time and effective cleaning takes more than two attempts every time. If it takes more than two attempts to effectively clean plant it is not financially effective and warrants considerable downtime, costs in lost processing opportunities, and the associated re-scheduling of production.

In summary, the survey results indicated a number of significant opportunities to improve cleaning. It was felt that the results from this survey indicated that the companies consulted did not have a true understanding of cleaning at the scientific level and were not able to identify contaminants. It was felt that more in depth information is required from industry to understand specific challenges. Industrial site visits were therefore recommended to increase understanding. This would also give firsthand knowledge of plant equipment, its geometry,

and organisation. Site visits to several manufacturing plants and the information obtained from the visits is discussed in section 3.3.

3.3 Information obtained from Industrial Visits with Britest member Companies

Section 3.2 showed that the information gained from the survey is valuable, but in order to gather more information it is considered vital at this point of the research to visit pharmaceutical and manufacturing plants to observe cleaning and increase the amount of knowledge known about plant cleaning. Site visits are considered invaluable as they are able to give a deeper understanding of specific challenges. It was possible to visit two companies with an interest in increasing their understanding of plant cleaning and discuss cleaning challenges. The information gained during the site visits is given in sections 3.3.1 and 3.3.2.

3.3.1 Site visit to Britest Member company 1

Company 1 is a globally integrated drug discovery company. It specializes in development, manufacturing and outsourcing solutions. Company 1 became involved with the plant cleaning project after realizing some of the benefits which occur with increased understanding of cleaning. Company 1 is a small multinational company who wish to examine a range of cleaning challenges surrounding several processes.

This company is unique among the Britest member companies as it has challenges in both food and pharmaceutical manufacturing. A lot of the problems in this company relate to equipment. A lot of large complex pieces of equipment including stirred tank reactors are used during processing and as it is a multi-product plant a lot of the equipment is used for more than one product. Process equipment used in one process (figure 3-14) shows the complexity of the equipment which adds to cleaning challenges. Company 1 has pieces of equipment that operators have been unable to clean effectively after processing particular products. The challenges associated with this have meant the dedication of plant equipment to products and in one case an entire section of processing equipment is not longer in use. The plant is also old in parts and this adds to the challenges, as the older the equipment becomes, the more damaged and worn it is. This requires operators to resort to manual cleaning more frequently and also carry out more dismantling of equipment.

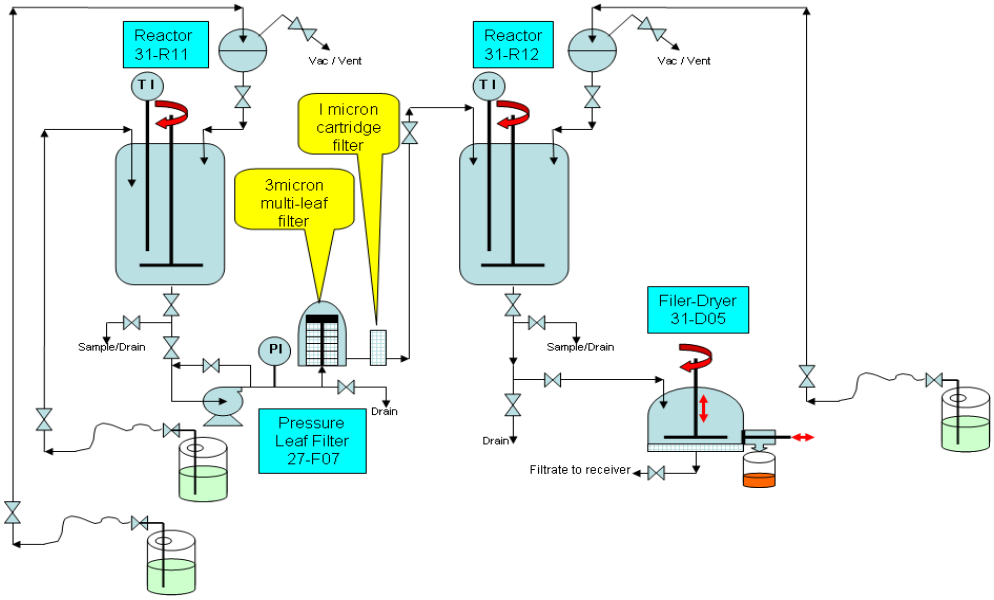


Figure 3-14 Equipment complexity as described by Company 1 (Company 1, 2011a)

In addition to other factors, this company were keen to stress during a site visit that the complexity of cleaning involves consideration of the materials of construction. This can include vessels, pipe work and gaskets, all of which can be in the same processing equipment and require cleaning together. Company 1 advised this is a critical consideration and should be considered when thinking of developing a cleaning model. Materials of construction can include glass-lined mild steel, borosilicate glass, stainless steel, Hastelloy and fluoropolymers such as Polytetrafluoroethylene (PTFE) or Edlon™.

Company 1 also discussed the recent adoption of improved cleaning equipment, including spray balls which follow a hydrodynamic spray pattern. This is beneficial as it provides better vessel coverage. It has been determined that stubborn contamination can be removed using the defined cleaning sequences.

Further cleaning challenges which Company 1 face are listed as -

Stubborn contamination which requires multiple applications of a defined cleaning sequence.

Highly coloured compounds (aryl azo compounds) visible at <1ppm and are therefore very difficult to remove from vessels and equipment.

Company 1 have a number of ‘sticky’ products which are food polymers and are very difficult to remove from equipment.

A major challenge for this company is sublimation of products and intermediate products into vent condensers/headers. Sublimated product is difficult to remove, due to the considerable complexity of this cleaning equipment and also because the contaminant is often unknown, and therefore the best method or solvent to clean the equipment is difficult to determine.

Another major challenge for this companies operators is cleaning 'Tide-line' residues on vessels (from degradation or evaporation). These are very common on vessels in most manufacturing sites. The manufacture of products is carried out to specific recipes or methods. This requires using the same volume of liquid in a vessel and therefore the vessel has a splash zone at the tide-line. This is depicted in figure 3-15.

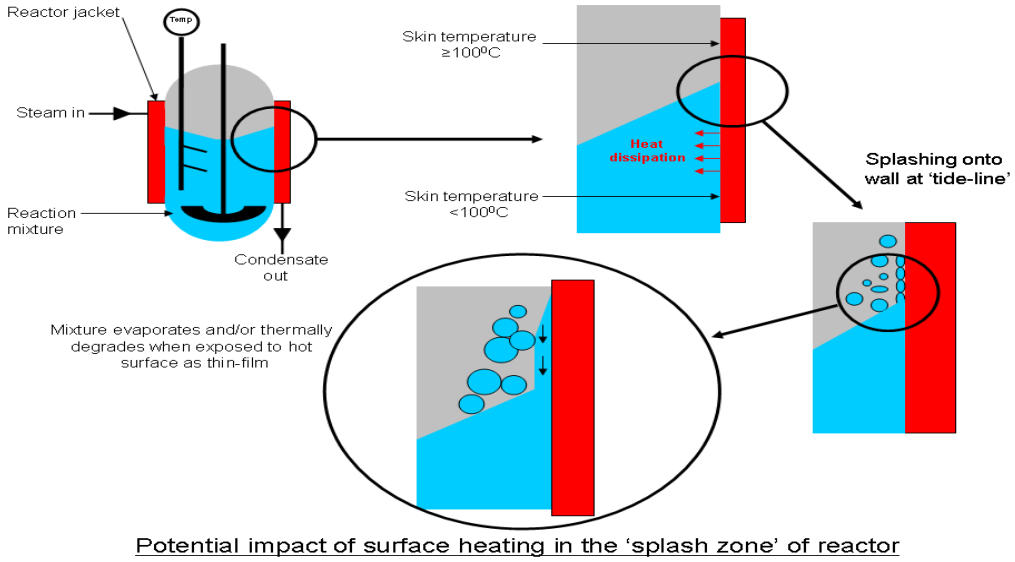


Figure 3-15 Splash zones in a reactor (Company 1, 2011b)

Mixed residues (organic/inorganic) can cause a considerable cleaning challenge, as it becomes difficult to determine which cleaning agent to use, and also to decide the order of use of any cleaning agents. This can lead to significant time lost due to trial and error.

Worn ('pitted') surfaces on vessels and other equipment are more difficult to clean.

Due to the complexity of plant equipment (figure 3-14), accessing some areas of plant to clean is difficult. As plant cleaning needs to be verified, inspected or occasionally swabbed to establish cleanliness, any inaccessible equipment becomes challenging to manage.

During the site visit to Company 1 it became apparent that determination of visually clean is difficult. Whether equipment is visually clean is often governed by who does the inspection. It is known that people have different standards and interpretations of vessel cleanliness.

The fact that many pieces of plant equipment have imperfections such as surface blemishes and scratches can compound the assessment of visually clean. It also means that any product or intermediate residue is likely to stick or gather in these areas.

Glass surfaces in reactors can suffer from markings or fogging caused by damage by water or chemicals. This is known as ‘Bloom’ marking. Bloom makes cleaning difficult as it cannot be removed as it is a change in the chemical structure of the glass .

Similarly ‘Rouge’ marking on stainless steel surfaces is a cleaning challenge.

The issues at this company’s plant have meant that management have determined that the general standard of cleaning needs improving. This is currently happening and standards are higher due to improvements in vessel cleaning and by increased understanding. This is being carried out using a contract cleaning company to determine the best use of detergents and solvents for cleaning. The American company Steris offer customised services which give solubility information in relation to their detergents and customers chemistries (Steris, 2011). This is good, but Steris do not divulge the nature or composition of their detergents. This is important as any product which is used in a pharmaceutical manufacturing process must be proven to be removed from the equipment, or present in low enough concentrations so as not to harm any recipients of any final drug product. If the composition of the detergents is not known how can their presence be detected, and how can they be proven eradicated from any equipment?

Company 1 raised a lot of cleaning issues which are useful to consider when developing cleaning tools. This company provided a lot of information and gave a good idea of their challenges which indicated that other similar companies may face the same difficulties.

3.3.2 Site visit to Britest Member company 2

Member company 2 is a large multinational pharmaceutical contract manufacturing company. The company is very interested in working towards improving cleaning standards and looking to reduce cleaning costs.

A visit to this company showed that the company is interested in reducing plant cleaning costs. This includes finding methods to ensure that the processes are carried out in the most sustainable manner, and continuously improving their processes to improve yields and reduce costs. The company operates to methods designed within the organisation that may not be the best industrial practices, but give an internal consistency to operations worldwide. The company also operates with ICH guidelines as the backbone for designing processes and methods and systems of operation, as is expected of a pharmaceutical company. The purpose of this site visit was to begin to understand cleaning challenges that the company face. It is a multipurpose site where the process equipment is in continual use for different products. This means that cleaning is critical and understanding it and improving it will help reduce plant down time and increase plant availability for processing.

At this site the manufacture of active pharmaceutical ingredients (API) is the main business concern. This means that the cleaning needs to be effective to reduce residual entities that could contaminate the next product into the equipment. This is especially important when manufacturing potent or toxic API's.

It is important to consider the level of cleaning required in these cases, which can change depending on what has been in the vessel, or what is made in the vessel next. For example, if the cleaning carried out in the vessel requires verification the maximum allowed carry over limit (MACO) may be higher than if the cleaning requires validation. The equipment cleaning required may be to validate between different products manufactured in the same vessel, or it may only need verification between batches of the same product in the same vessel. The protocols or procedures for cleaning may therefore change depending on the standard of clean required. If cleaning verification is required a short cleaning protocol or part of a protocol may be used. If cleaning validation is needed the full cleaning protocol may be used. This site visit was useful at this stage in the research to determine the importance of plant cleaning. It has also been useful to help determine answers for three of the research questions asked at the beginning of Chapter 3. This is discussed in the next section 3.4.

3.4 Research Question answers

At the beginning of Chapter 3 three research questions were asked.

RQ3: What are the main challenges associated with process plant cleaning for Britest members?

RQ4: What common methods do Britest members utilise to clean their process equipment?

RQ5: Which cleaning challenges are associated with plant engineering or choice of cleaning agent?

Through the answers received from the survey it is possible to answer these questions.

RQ3 Answer: The main challenges associated with process plant cleaning for Britest members are achieving right first time cleaning and understanding the fundamental science behind the cleaning. The fundamental science behind cleaning is not known by Britest members although the contaminants may be known in some cases. The cleaning seems to be based wholly on solubility with little regard to deeper understanding (Figure 3-9). This approach to cleaning is challenging, especially because cleaning is not considered early enough in the manufacturing process to fully understand the science behind cleaning.

RQ4 Answer: The common methods used by Britest members to clean their plant equipment are targeting a specific area for cleaning with a specific cleaning agent, flushing an area with cleaning agent and isolation of pieces of plant equipment for cleaning. This may involve disassembly of equipment to ensure that it is cleaned efficiently. In addition the survey results indicate that the cleaning agents commonly used are organic solvents, mineral acid or alkali. As well as solvent cleaning a majority of respondents used water and aqueous detergents.

RQ5 Answer: The answers to **RQ5** are strongly linked to the answer to **RQ3** and **RQ4.** The challenges associated with plant engineering and cleaning agents are achieving right first time cleaning. This is the most favourable cleaning situation. In addition, time spent isolating, dissembling equipment and waiting for it to be confirmed clean are challenges associated with a majority of Britest members. This is very time consuming and increases the amount of time a piece of plant equipment cannot be used for manufacturing.

The survey results show that there are common challenges for Britest members. It is important at this stage of the research to examine the suite of tools and methodologies which are within the Britest portfolio. This will help to shape the research and begin to start to address the cleaning challenges by finding a tool or methodology which can be used to aid Britest members in cleaning their plant equipment. Firstly, in section 3.5.1 Britest tools are introduced. In section 3.5.2 Britest tools will be examined further, by determining which tools can be used by Britest members to help solve cleaning challenges.

3.5 Britest, Britest Tools and Methodology

3.5.1 Britest

The Industrial sponsor for this project on industrial cleaning is Britest. Britest is a not for profit company which is owns intellectual property in the tools and methodologies, which Britest members have access to. In addition a related benefit is that they provide an area for members to share best practise at a pre-competitive level. Britest is a way for members to collaborate to develop solutions to key common challenges in order to drive knowledge forward. It is stated that, '*member companies benefit from access to a range of propriety tools and methodologies focused on whole process understanding*' (Britest, 2011). It is estimated that Britest have delivered over £500 million of tangible value to their members to date (Britest, 2011). Benefits from using the tools include higher product yield, enhanced product quality and, critically, a robust and understood process. It is hoped that through this project, tools can be developed to improve the understanding of industrial cleaning and improve sustainability of day to day plant operations. The development of appropriate tools for these areas would complement the existing tools, bringing whole process understanding to another level.

3.5.2 Tools and Methodologies

Britest have a portfolio of tools and methodologies which are successfully used in manufacturing sites throughout the world to identify and solve process related challenges. Britest tools encompass a range of qualitative and semi-quantitative models. The models allow users a method with which to capture knowledge and critical information in relation to their processes, in a way that can be easily understood and further analysed. Importantly, the models also highlight lack of knowledge and understanding in some areas. The methods allow the users to consider information from other experts within their organisation as multidisciplinary teams work together to resolve issues. In addition, Britest team members work alongside the member teams, teaching them how to apply the tools through training, and bringing their previous experience and skills to facilitate their use. The Britest tools provide an effective way to focus teams and improve communication to reach solutions and solve problems. Importantly, this can only work if the team members involved are the right people, with the correct process understanding, and are allocated time and space to work through the process using the models.

Britest methodologies are described as the 'expertise and knowledge required to obtain the best answers by using the correct tools'. This is important as only the correct tools will

provide the best solutions to problematic situations. The methodologies allow consideration of the best tool for the situation to be chosen. They allow logical and focussed thinking around an issue.

The methodologies also allow Britest members to get the best value out of the tools, ensuring they are used in the correct manner to obtain better process understanding. This may be by Britest Lite, or by a full Britest study. A full Britest study is carried out by a number of experts in the area of consideration. For a process related investigation this may include process operators, analytical staff, chemical engineers and chemists. Everyone who is important to aiding the production of a solution should be part of the study. The full study takes a day or longer and will be in-depth, focussing on individual aspects of chemistry and engineering that may be altered to provide a possible solution. The application of most Britest tools will be considered in order to reach the best potential outcome.

It may be that a very specific plant cleaning problem can quickly be investigated and potential solutions identified using Britest Lite. Britest Lite is a shorter condensed version of a full Britest study. It is useful if less time and people are available to participate in the study. Britest Lite is used for problem solving in specific areas and fewer tools will be used in this study. In a potential plant cleaning study of this nature it may be possible to bring a small group of knowledgeable people together to solve a specific problem with Britest Lite.

These kinds of problem solving can be carried out in house, after initial Britest enablers have been put in place. Britest enablers are methods of introducing and empowering the Britest members to problem solve using the Britest tools and methodologies. These include training packages in specific Britest tools, website resources and background knowledge that allow the best application of the tools. Tool selection is considered an important aspect of the Britest methodology. There is no reason why a number of tools cannot be tried in order to achieve the best fit tool. A vital key to the use of Britest tools is realising what the challenge is, and which question requires an answer.

Core methodologies of potential interest in this research include the following; Initial Screening Analysis (ISA) and Duty Definition and Equipment specification (DuDEs). ISA is useful methodology that can be used to start any Britest study or Britest Lite study. ISA is a method that gives an overview of the whole problem through a number of stages.

Initially the methodology tool known as ISA is used to define the problem, so that the team knows what the outcome of the study needs to be. This means that the purpose of the study is known, such as reducing the cost of a cleaning step by using less resources.

Once the purpose of the study is known, the scope of the study can be considered. An example of this is a plant cleaning problem in a specific process, but the study is constrained to finding a solution for one part of the process. This is useful as it brings a focus to the study and it is considered that this tool will be important in future plant cleaning investigations.

The intent of the study is then to define the product, which may be a pharmaceutical or chemical product. This is critical as understanding the product and the product characteristics indicates how the product interacts and the process reactions occur. The product required can be specified in terms of its physical and chemical properties. Once the product is defined then the key by products and reagents can be understood and identified. Therefore the process used to make the product can be defined. The identification of further tools and methodologies required to best indicate potential solutions to the problems can now be addressed.

Duty Definition and Equipment specification (DuDEs) is a methodology that may be applied in future plant cleaning case studies. It is used primarily for process decision making, such as whether to buy new equipment or for justifying capital investment. Both of these scenarios could be possibilities for case study companies during the plant cleaning project. It may be that a solution to the cleaning problem is to buy a specific CIP unit. This would require a tool such as DuDEs to help select the correct equipment or identify modifications to existing plant. This methodology has the following stages.

1 Initially, process understanding would be clarified. This is important to the plant cleaning project, as the fundamental chemistry and physics behind the process need to be identified and appreciated.

2 A process plan would then be created that allowed all processing options to be considered carefully.

3 Equipment specification would then be established to determine what the new or modified equipment needs to be able to do.

4 Research and analysis of the equipment would then be carried out to select the best option. This would be carefully reviewed prior to proceeding to the next stage.

5 After these stages, the equipment may be selected and designed.

Further Britest methodologies consist of the following tools and models. An indication of the application of the tools to plant cleaning is now considered.

Process summary information map (PrISM) is a tool which is incorporated into the ISA methodology. It is specifically used to represent a process in a map format. It shows the main stages in a process and therefore focuses attention on known facts that happen during processing, such as what are the process inputs and outputs. It can also be used to capture known relevant information, including processing costs at each stage. During plant cleaning research this tool may be used to break down process stages into smaller steps to begin to understand chemical interactions and the formation of by-products and intermediates. It is thought that many of the by-products could be a cause of difficult to remove contaminants in Britest member's processes. This is therefore a good tool to capture and analyse known information and it may be critical in discovering gaps in process knowledge.

Process Definition Diagrams (PDD) is a tool that uses knowledge from chemists and chemical engineering and brings them together in one place to consider the effect that they have on each other and on the overall process. The tool uses a form of state – task approach. This means the process is described as a series of tasks which transform starting materials into products. This is a fundamental concept of the plant cleaning project. The information is collected in a way that describes the process, in its simplest form. This means that factors such as scale and the equipment used are not considered. The plant cleaning project will focus on chemical, physics and equipment interactions. It is therefore considered that PDD could be used with another tool or adapted to encompass this. The mechanisms for undertaking this are yet to be determined but it is considered that the flow of a cleaning agent could be tracked through plant equipment in a similar way to a product to determine what happens to it. A simple method for including this information may be to annotate the PDD according to the equipment used, by the addition of a series of developed symbols.

Rich Pictures (RP) is a simple but powerful tool. It enables a clear description of a process stage or piece of equipment, and can help determine what is physically happening in a great amount of detail. It can be applied to chemical reactions and fluid circulation which is important in plant cleaning. This has been used in previous cleaning studies to determine areas that a spray ball could not reach inside a vessel. It is therefore thought that this technique will be of importance during the plant cleaning project. The visualisation of a vessel or a piece of equipment can indicate many factors that have not been considered previously and therefore can suggest solutions to problems. It is thought that a number of RPs could be

used for varying flow rates of cleaning fluid, or positions of jet washers of spray balls to determine the scope of their influence. This would need to be carried out in conjunction with analytical methods that indicate clean and unclean areas. These could then be mapped onto the RP and analysed to determine potential reasons for differences. Figures 3-16 and 3-17 give examples of how RP may be used to model complex process tasks by showing coffee making in a cafetiere.

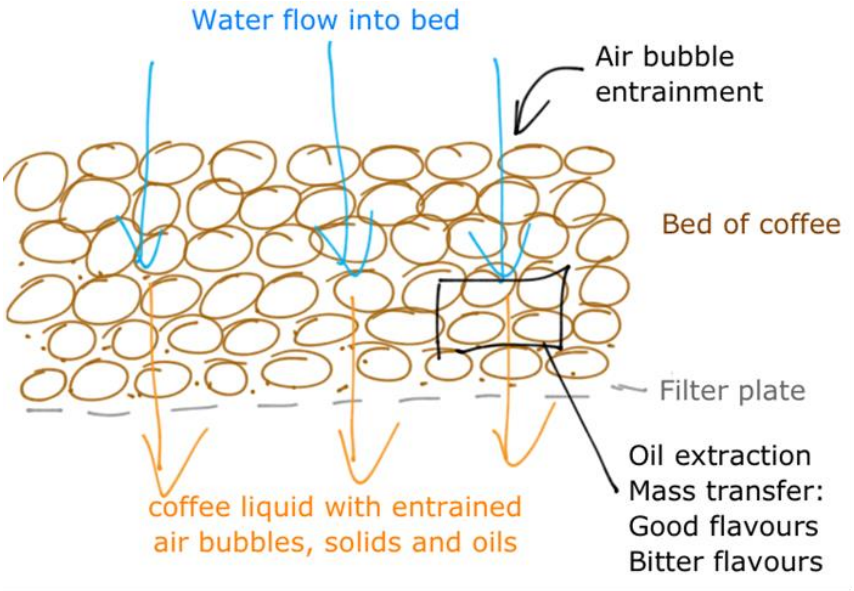


Figure 3-16 Inside the cafetiere where water is poured onto a bed of ground coffee beans. The Rich Picture is able to capture and visualise changes taking place in the cafetiere (Britest 2016).

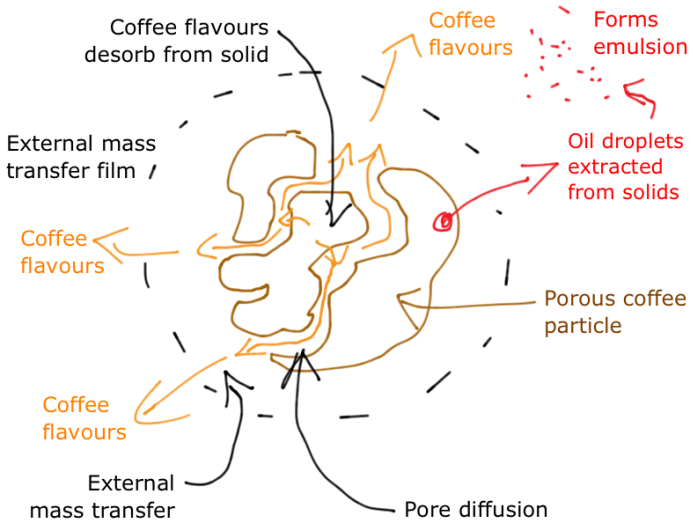


Figure 3-17 Detailed RP giving finer detail about how the coffee is made and the considerations and ideas which could arise from a discussion around making coffee in a cafetiere. (Britest, 2016)

Rich Cartoons (RC) are similar to RP and could also be used effectively during the plant cleaning project. They allow visualisation pictorially of movement through stages of processes with time. A cartoon drawn of a process stage can show how chemicals interact and raise questions about what potentially remains at the end of a stage. It can be drawn to reveal limitations and solutions within processes. Ultimately it shows how events unfold in what sequence they occur and what is happening at the same time. It may be apparent from this that factors not thought to interact do interact. This may be very valuable when considering the nature of contamination and how it occurs. RC could potentially show that contaminants are produced as a result of parallel reactions not previously considered.

Transformation Maps (TM) are used to show how the sequence of rate processes, both chemical and physical rate processes happen. TMs show reactions can produce wanted and unwanted products. It can be used to distinguish intermediates and also show how reactions happen. An example of a TM for aspirin indicates where all of the reactions take place and where phase changes can occur (Figure 3-18).

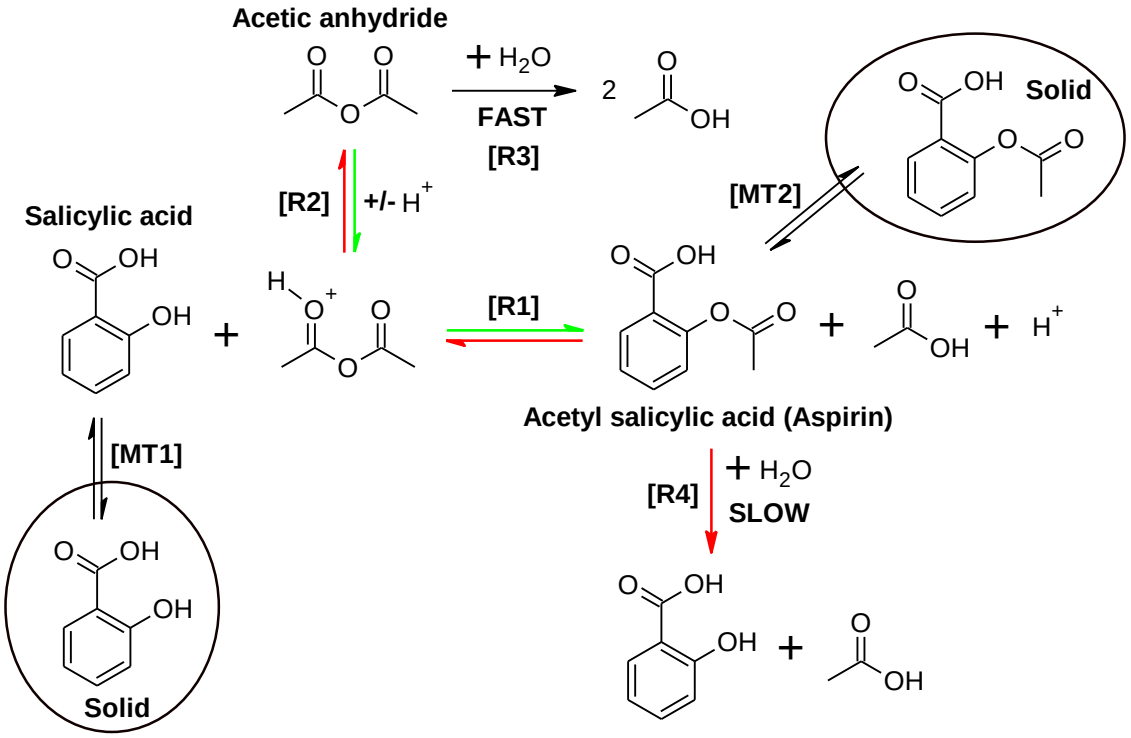


Figure 3-18 Transformation Map (TM) of Aspirin (Britest, 2016)

Notes for Figure 3-18:

Where MT is Mass Transfer and R is Reaction

Where arrows are coloured, green indicates a desired reaction or direction of equilibrium, and red indicates an undesired reaction or direction of equilibrium.

Figure 3-18 gives a good indication of the level of detail which can be achieved using a simple product TM. The above TM shows states which form during the process and can indicate a good cleaning solvent for used equipment. It is known that the solubility of both acetyl salicylic acid (3 g/L at 20 ^{o}c) and salicylic acid (2 g/L at 20 ^{o}c) is relatively low. This can often cause the product to contain un-reacted salicylic acid. This needs to be removed by re-crystallisation. This implies that water would not be a good choice of cleaning agent for cleaning manufacturing or plant equipment used during this process.

TM is often used as a tool to clarify information prior to using Driving Force Analysis (DFA). TM will be of use during this project to potentially identify contaminants and hopefully it can be used in conjunction with DFA to prevent or limit their formation.

DFA is a tool that identifies in the form of a matrix which factors influence the rate of product and side product formation and the rate of formation. It can be very powerful in determining if the factors that influence reactions truly reflect observations during plant operation. This could be used to reduce or eliminate contaminants in processes by altering factors that can limit or drive their reactions. Different chemistries or conditions can then be utalised as part of whole process design.

Transformation, Entities, Properties, Physics, Parameters, Order of magnitude (TE3PO) is also thought to be useful for the plant cleaning project. This tool enables teams to focus on physical processing and transformations. It identifies in a structured manner the key entities and properties in a transformation as well as the fundamental parameters that govern the transformation. Chemical processes can be complex and this gives a logical approach to thinking through those processes, with regard to the entities that are present in the process stage, what the physical properties might be and what physics is needed to go through the required transformation. All these factors are important when considering how to remove contaminants from processes. TE3PO could therefore indicate practical methods of contamination removal during this project.

The application of the tools is dependent on the member companies' requirements and can be applied to the whole process or to a specific area of the process. Two case studies will be discussed in this report which successfully used Britest tools to examine process challenges. It is thought that initial plant cleaning study tools may include ISA and PDD; it is considered that a combination of RC and RP will be useful during PDD assessment when considering staining or residue analysis. Depending upon the process study, it may be useful to carry out

TM and DFA to indicate how to prevent the formation of contaminants.TE3PO may provide useful information on how to remove contaminants.

It is known that the established Britest tools may not provide solutions for all cleaning problems encountered. It is thought that the influence of the type of equipment and scale of the process is not truly explored in the current tools. This was developed during the course of this research as it was anticipated that the equipment type contaminated and CIP used were factors influencing removal of contaminants. This appeared to lie beyond the scope of the current tools, with the exceptions of RP and RC.

Examination of the current Britest tools and methodologies has indicated that there is scope to use them to develop a suite of tools specifically for cleaning. The current Britest toolkit does lack a tool to help choose a cleaning agent or methodology early in process development, and it is this tool or methodology which this research must focus on. Focussing on this issue will mean right first time cleaning occurs more frequently than currently stated in the survey in figure 3-13. One fundamental question which remains unanswered in this research is, how much does failing to clean right first time cost? In order to answer this question it is critical to examine the costs or metrics associated with plant cleaning, and determine if current plant cleaning is cost effective. This will be considered in section 3.6.

3.6 Plant Cleaning Metrics

Research information from AstraZeneca suggests that large pharmaceutical companies can spend up to 50% of plant time carrying out cleaning. The average time associated with this is 500 man hours and 8 tonnes of solvent with an associated cost (including downtime, labour, waste treatment and lost opportunity) of £1 million per process clean (AstraZeneca, 2008). Smaller companies involved in fine chemical manufacture indicate that the downtime associated with cleaning can be up to 20% of plant time. The cost of the 20% of unused plant capacity is £20 million. This equated to approximately 10% of turnover (Britest, 2009).

Cost Benefit Analysis (CBA) is important in industry, including the pharmaceutical sector. This is because in order to justify changes in any process or operations it is necessary to determine financial impact. The industrial plant cleaning project looks to reduce plant cleaning costs and therefore it is useful to consider applying this analysis to cleaning in the pharmaceutical industry.

In order to fulfil this remit for the research project there must be an improvement in the cleaning methods or techniques applied. This must be measurable either by financial or other

means. It could be considered that the metrics for processing pharmaceuticals are well documented. This is not the case. It is very difficult to gain access to financial information from companies who consider that sharing this information may be disadvantageous for a number of reasons, for example, giving other companies a competitive advantage if solvent purchase prices are shared publicly. It is also apparent from information gained in the survey discussed in section 3.2, that companies who took the survey do not break down operation costs for every process in terms of overheads. Although costs of API's are known and solvent prices are known the cost of the operation of specific equipment is not often considered. This makes it difficult to analyse the costs for overall pharmaceutical processes and more difficult to determine the costs of cleaning equipment.

For the purposes of this research the cost of specific pharmaceutical products can be found in the Drug Tariff. For example, one of the most expensive drug products on the market in the UK is Pramipexole (3.15mg modified release tablets) at £12.03 per tablet. One of the cheapest drugs on the market in the UK is Aspirin at approximately 4 pence per tablet (NHS, 2013). The cost of an off-patent drug is based on the cost of the process, the cost of the raw materials and the size of the batch. For drugs that are still under patent a significant cost is the markup for the pharmaceutical company which aims to recover the cost of research and development costs which can be considerable. One of the biggest costs during any manufacturing process is often the raw materials. The cost of solvents fluctuates but some can be significantly more expensive to buy, to use, and to dispose of (as discussed in section 3.6.1). It is therefore critical that solvent use is kept as low as possible in order to reduce costs. Costs associated with solvents include purchase, storage of solvents pre and post use, and disposal of solvents. It is clear from the industrial information provided in section 3.2 that this is not always possible. This is due to repeat cleaning costs, the inability to achieve cleaning RFT. These factors and plant downtime are potentially not factored into the cost of drug production. In order to achieve this a lot more information must be understood about cleaning metrics.

Cleaning metrics for aqueous cleaned systems in the food and cosmetic industry have been determined by Benson (2009) with the use of a benchmarking tool.

In order to determine the cost of aqueous cleaning Benson, along with Ecolab Ltd, developed a tool which could easily and quickly calculate the cost of cleaning using non solvent cleaning agents. This work was sponsored by the Technology Strategy Board in 2009. The resulting tool was called Zero Emissions through Advanced cLeaning (ZEAL). It was considered important to examine modifying this tool to gain information on cleaning metrics for the use

of solvent based cleaning. One of the Britest members involved in the survey (section 3.2), Company 3, was chosen to participate in this exercise. The industrial plant cleaning project looks to reduce plant cleaning costs and therefore it is useful to consider applying the Zeal database to the pharmaceutical industry. The Zeal database has been used to great effect by many companies who wanted to assess and analyse cleaning methods. The tool provides a valuable method for collation of data and allows identification of areas for improvement, determining where money, time and resources could be reduced if changes are made.

An on-site visit was carried out with Company 3, who had an interest in reducing plant cleaning costs in order to begin to understand how the ZEAL tool could be adapted to provide metrics for this research. Company 3's interests include finding methods to ensure that cleaning processes are carried out in the most sustainable manner, and continuously improving cleaning processes to improve yields and reduce costs. The company operates to methods designed within the organisation that may not be the best industrial practices, but give an internal consistency to operations worldwide. Company 3 operates with ICH guidelines (Chapter 2 Section 2.2.7) as the backbone for designing processes and methods and systems of operation, as is expected of a pharmaceutical company. The purpose of this site visit was to begin to understand cleaning challenges that the company face on site. It is a multipurpose site, where the process equipment is in continual use for different customer products. This means that understanding and improving plant cleaning is critical, and will help to reduce plant down time and increase plant availability for processing.

At the site the manufacture of active pharmaceutical ingredients (API) in a multipurpose plant is the main business concern. This means that the cleaning needs to be effective to reduce residual entities that could contaminate the next product into the equipment. This is especially important when manufacturing potent or toxic API's.

It is important to consider the level of cleaning required in these cases, which can change depending on what has been in the vessel, or what is made in the vessel next. For example, if the cleaning carried out in the vessel requires verification, only the maximum allowed carry over limit (MACO) may be higher than if the cleaning requires validation. The equipment cleaning required may need to be validated between different products in the same vessels, or it may only need verification between batches of the same product in the same vessel. The protocols or procedures for cleaning may therefore change, depending on the standard of clean required. If cleaning verification is required a short protocol or part of a protocol may be used. If cleaning verification is needed the full cleaning protocol may be used.

This information is not currently captured on the ZEAL database, as cleaning in the food and (dairy industry) and the cosmetic industry does not require similar levels of cleaning in terms of verification and validation. In order to make the database suitable for the pharmaceutical industry, this information needs to be considered and captured to reflect its importance.

In addition to this, further site visits highlighted other information and differences between the ZEAL database, and the data it suggests contribute to cleaning costs and the requirements of the pharmaceutical industry. These are listed and discussed in the sections indicated:

- Waste product differences leading to waste treatment diversification of cleaning waste treatment and cost of waste disposal (Section 3.6.1).
- Cleaning standards verification and validation (Section 3.6.2).
- Analytical methods and time taken to analyse samples (Section 3.6.3).
- Multi process operation by staff (Section 3.6.4).
- Multi produce use compared with the dairy industry or the brewing industry (3.6.5).
- Additional differences and information (3.6.6).

3.6.1 Waste Disposal

One of the main differences between the food industry and the pharmaceutical industry is cleaning waste disposal. In the food and cosmetic industries cleaning waste disposal is generally considered easier. Cleaning waste generated in the pharmaceutical industries is more complex and can include a lot of solvents used in cleaning (and during processing) and also water contaminated with solvent and other residual elements, some of which may be toxic and require specialist disposal.

This means that companies who manufacture pharmaceuticals may have several complex disposal routes for each manufactured product. An example of this is product X manufactured by company 3. The waste disposal route required for the cleaning products is shown in figure 3-19.

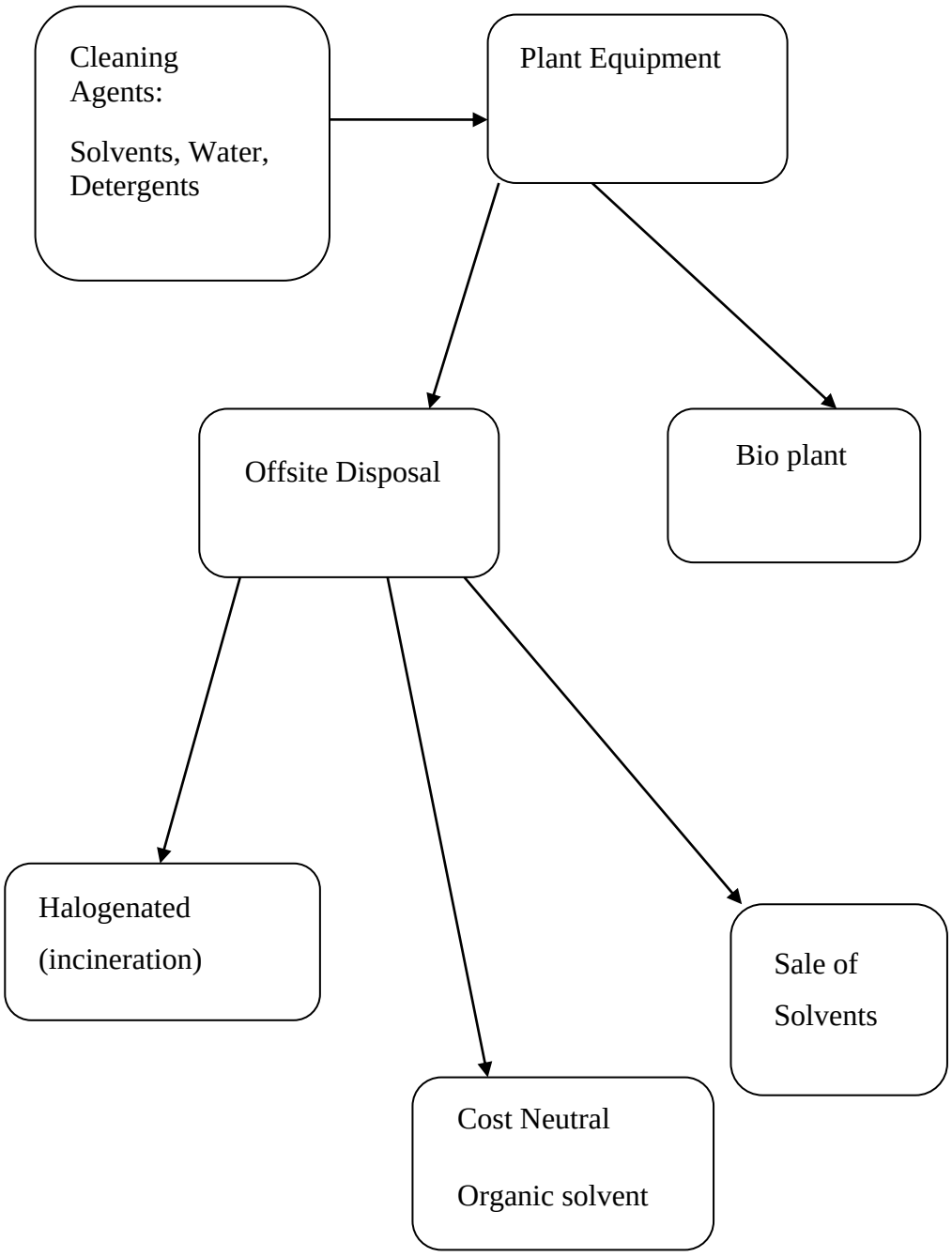


Figure 3-19 Waste Disposal from Process X at Company 3 (Generated from information provided by company 3).

As figure 3-19 indicates waste disposal can be complex. Waste is either disposed via the bio plant or it is disposed of offsite, which is more expensive as it has additional transport costs associated. For this particular process X there are three waste disposal routes. The first

concentrates on disposal of halogenated waste, which requires expensive treatment by incineration. The second route is deemed cost neutral and disposes organic solvent. The third route involves the sale of some solvents which generates money for this company.

It is possible to say that the above figure is complex, but in addition to this the costs incurred in packaging the waste also need to be considered in the ZEAL database. The waste disposal on site is more cost effective to dispose of, as it can be directly pumped to the bio-plant. Costs increase when the waste disposal is carried out off site.

The initial cost of waste disposal is determined by how the waste is packaged. There are several options available for this. The waste may be packed into 500 or 1000L intermediate bulk containers (IBC's). This is more cost effective than other methods but it cannot be used for all types of waste. This is due to IBC materials of construction; some solvents are not suitable for use with these vessels.

A more expensive method for disposal of waste is to put it into 200L barrels. The challenge associated with this method is that the barrels are expensive to loan or buy. Once barrels are used they are returned to the company where they came from. The barrels may not be clean when they arrive on site and therefore any waste solvent deemed saleable must be put into clean or new barrels so the contents do not become contaminated by the barrels. Only solvent content above a certain limit can be bought by companies, for example for use in car windscreen wash. This requires careful operation of cleaning waste and knowledge of the barrel contents post use. Another issue with barrel use becomes apparent when halogenated waste requires disposal. This is expensive to carry out. In these cases the barrels are incinerated and cannot be reused or returned to the company, which increases the cost of disposal by this method. Another factor which requires consideration is whether the waste is stored on site and the associated cost this incurs until the waste can be disposed of. Overall, packaging, disposal and potential storage of waste can greatly contribute to cleaning costs, which is not considered in the original ZEAL database.

The addition of this information to the ZEAL database for pharmaceutical companies would show how much was spent on waste disposal, and this would lead to suggestions and options to reduce these costs. Waste contents are important, as this determines how it is disposed and how much it costs to do this. Cleaning standards are also important and can impact upon the financial cost of cleaning.

3.6.2 Cleaning standards verification and validation

Unlike the food and cosmetic industries, pharmaceutical companies operate to cGMP standards and within regulatory guidelines. This has already been discussed in Chapter 2 sections 2.26 and 2.2.7. The food and cosmetic industries have their own regulations and operating regulations. In the food industry for example, identifying hazards and critical control points (HACCP) is applied. In the food and cosmetic industries processing equipment is kept clean, but there may be only one type of cleaning. When cleaning is carried out in any industry, the equipment is either clean, or not clean. Generally clean is determined by visible inspection or by assay (Refer to Chapter 2 section 2.2.5). In the pharmaceutical industry cleaning can either be verified or validated. Importantly, survey results in section 3.2 (Carr, 2011) indicated that a number of Britest member pharmaceutical plants operate as multiproduct facilities. This may not be the case in the dairy industry, which the database was designed for.

This information is not currently recorded on the ZEAL database, and it is an important factor in pharmaceutical cleaning. A verified clean may only require cleaning equipment with part of a cleaning protocol, or omit dismantling equipment. This not only saves time but means that cleaning could take less manpower, therefore needs to be captured on the ZEAL database. Equipment dismantling cleaning and reassemble is time consuming for Britest members.

Validated cleaning is stringently controlled cleaning and the method for validation can vary between companies. This needs to be captured on the database to ensure that time and resources can be properly calculated for these operations.

One of the most important omissions required for recording pharmaceutical cleaning is the number of cleans taken to achieve the required standards (either verified or validated cleans). Although the 2011 cleaning survey (section 3.2) showed Britest members were confident enough to state their cleaning protocols achieved good consistent results, speaking to members showed this was not true. In reality cleaning in many companies does not achieve the required results every time and cleaning may need repeating. The database must be adapted in order to reflect this, as if a vessel is not cleaned right first time it costs more money, time and resources. This can have a negative effect on processing and delay it significantly. An example of a vessel not cleaned right first time is shown in figure 3-20.

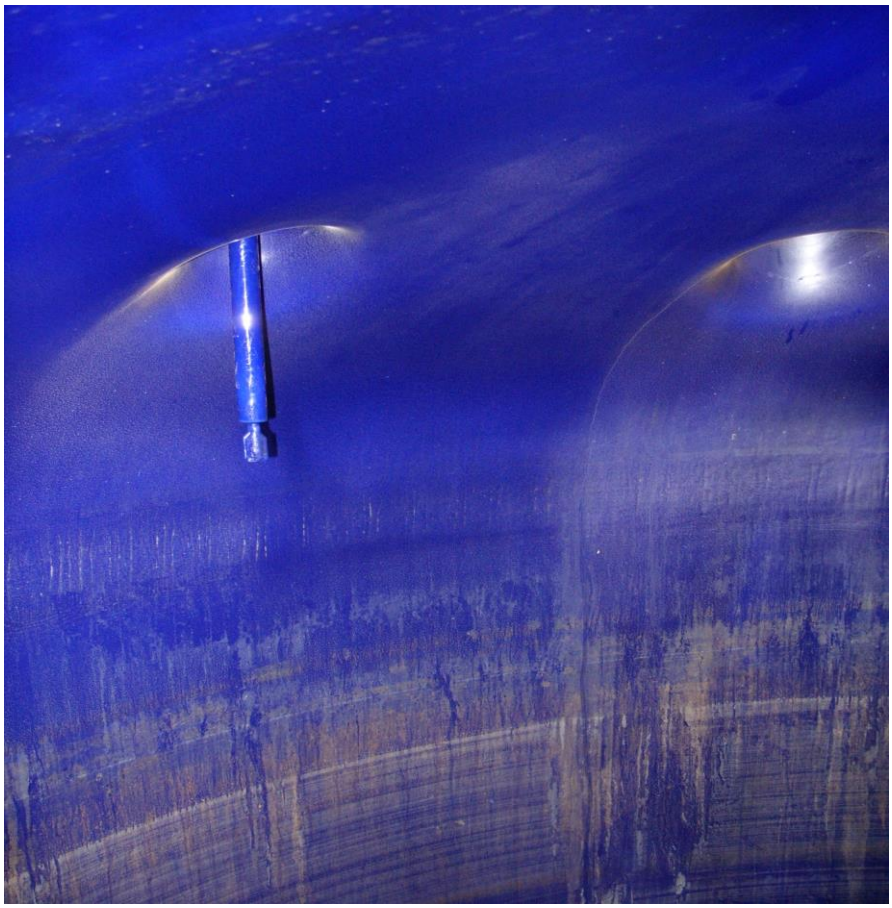


Figure 3-20 Example of Stained Blue Glass lined vessel post cleaning (Company 3, 2012).

Figure 3-20 shows vessels are not always cleaned right first time. A key achievement linked to this project is to help achieve right first time cleaning. There is a lot of staining and residue left on the vessel post cleaning. Some of this residue is clearly apparent as a tide mark line, which corresponds to cleaning challenges indicated at other surveyed companies, including company 1. It is obvious if a visual inspection is carried out on a vessel and it appears as above that cleaning will have to be repeated. At this stage analytical methods would not be carried out. Analytical methods will be discussed in section 3.6.3.

3.6.3 Analytical methods and sample analysis time

The analytical methods and time taken to analyse samples is not recorded on the ZEAL database. This is due to the amount of time it can take for assay results to become available. Often in industry cleaning assays can take time to carry out, therefore results can take time to reach production managers. In order to continue manufacturing in the same vessel, clearance by quality control and quality assurance is needed. When this occurs is dependent on many

factors. This can include the time quality control staff are available to give the equipment clearance. Sometimes communication is poor between departments and critical assay results need to be chased. Assay priorities may not fall with cleaning assays. Final product assays take priority in a majority of cases, delaying cleaning results. It is recommended that the ZEAL database is modified to incorporate this information.

The failure of cleaning assays or assay repeats are commonplace in industry for many companies. If this does occur equipment downtime can increase while awaiting results.

If equipment is not required for processing post use, assay results can be delayed further while other assays take priority.

Another key factor not represented on the ZEAL database regarding cleaning analysis is drying time. In order to properly assess vessels visually and also take swab samples, vessels need to be dry. Drying time can be very variable between vessels and the same vessel at different times of the year. This information needs to be collected in order to accurately assess how long equipment remains out of use. This is particularly relevant when considering the role of operators on plant as discussed in section 3.6.4.

3.6.4 Multi process operation by staff

Due to current manufacturing processes it is possible for more than one process to be operated by one person. This makes it difficult to apply costing for a labour resource to cleaning, particularly when waiting for cleaning cycles to finish and equipment to dry. In order to achieve this, careful consideration must be given not to overestimate labour resources applied to cleaning. This must be captured onto the ZEAL database.

3.6.5 Multi produce use and Product Types

Product types require alteration for the pharmaceutical industry. This is due to two reasons, which are firstly, that Pharmaceutical plants tend to be multipurpose, based on the evidence gained from the survey results in section 3.2. Secondly, definitions need to be altered from dairy milk and cream (fat, protein and carbohydrate) to appropriate products and residual types for the pharmaceutical industry.

3.6.6 Further Database Adaptations

In addition to the above adaptations there are several software changes which could be carried out in order to make the database more applicable for the pharmaceutical industry. The key assessment information required for cost benefit analysis is listed below indicating adaptations to each sector.

I - Utilities, labour, effluent and product and chemical costs

Cleaning waste such as packaging is not recorded in the ZEAL database. It is recommended that this is included in the database, as it represents significant cost to pharmaceutical companies disposing of solvent waste.

The number of hours of labour or man hours is not recorded for people who carryout multiple tasks at once. It is recommended that the database should be altered to allow this to be captured.

It was discovered that changes to recording cleaning waste are required in the database, including packaging waste. In addition, labour use (in person hours) should be recorded with regard to multi-process operations by employees. That is, if people are carrying out multiple tasks, then the proportion of time on each process task should be recorded.

II - Cleaning Scenario

Product types need adaptation to the pharmaceutical industry.

III - Cleaning Times and Consumptions

The ZEAL database should be adapted to account for disassembly of equipment and equipment drying time. Any time taken to assay and feedback the assay results, and clear the cleaned vessels for use is not captured on the database and should therefore be included. The type of cleaning required either verified or validated should be incorporated in this section.

IV - Site information

No adaptations.

V - Cleaning Data

Addition of analytical information is required.

VI - Process and Cleaning Diagrams

Ensure incorporation of disassembly information and transfer times taken to do this onto cleaning times.

VII - CIP and effluent monitoring

Capture internal and off plant waste disposal and packaging waste.

VIII - Management View

No adaptations.

IX - Overall Cost Summary

Ensure that the cost summary reflects changes and adaptations to give the best representation of cleaning costs.

The above adaptations would help capture cleaning information and therefore allow cost based analysis. This would result in improvements based on changes. This was taken into account when using the ZEAL database with Company 3, as described in section 3.7.

3.7 Cleaning Cost Benefit Analysis for Company 3 using ZEAL database

In order to determine the cost of cleaning for Company 3, one post process cleaning process was examined. Examining the process highlighted a lot of differences between pharmaceutical and the industries the database has been previously used with, as discussed in section 3.6.

Although Company 3 provided a lot of information to enable cost benefit analysis of one cleaning process, more information needs to be provided in order to fill in the database. It was determined the best method to carry this out was to analyse the cleaning process by using the Britest tool Process Definition Diagrams (PDD). Using this technique it is possible to reveal what each vessel contained at each stage of the cleaning process, and what the conditions were inside the vessel in terms of temperature and holding time.

This highlighted other issues with the cleaning process in relation to the information shown on the PDD. In order to clean vessels and pipes, it is important to consider what residues and potential contaminants are present, which the PDD can do, but, it is also important to consider the age of the vessels and the materials of construct and geometry, which can affect cleaning. This is not generally shown on a PDD. Therefore the PDD tool itself needs adapting to this purpose. The PDD model will be adapted to a Process Definition Cleaning Diagram (PDCD).This is due to the fact that cleaning is a process that relies upon the correct treatment of multiple variables which affect it. The vessel geometry, age and material of construct have important roles to play in determining the effectiveness of a clean. This may be the reason why cleaning produces variable results and does not always result in right first time cleaning scenarios. A standard PDD is shown in figure 3-21.

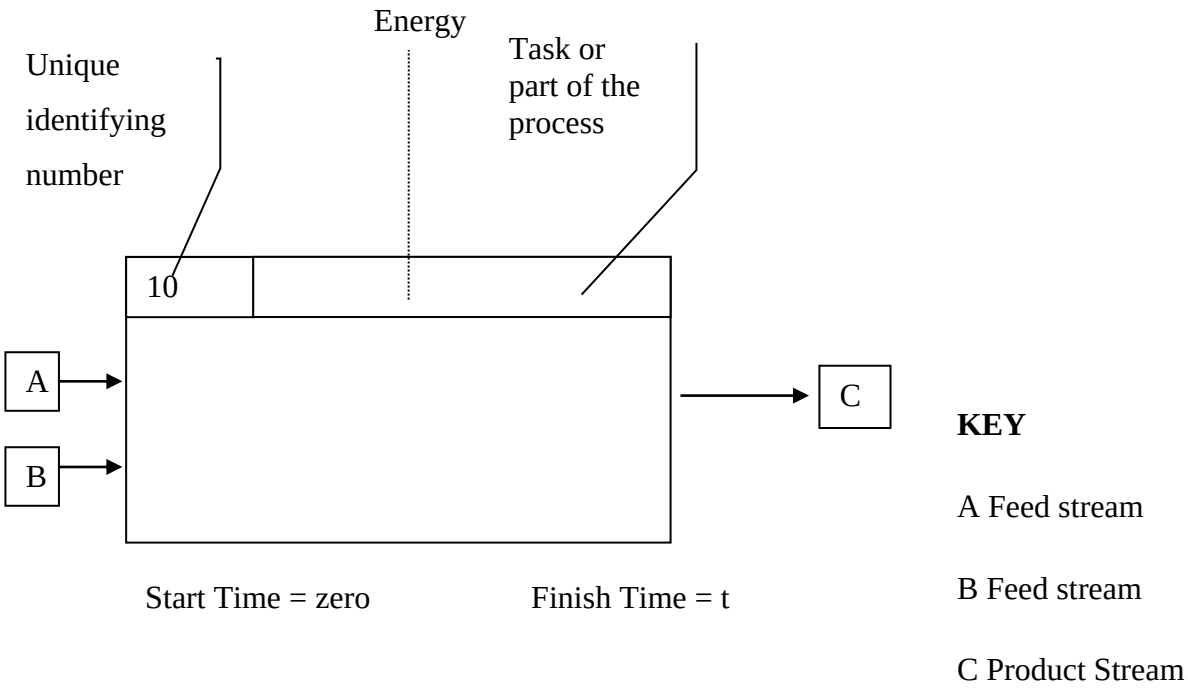
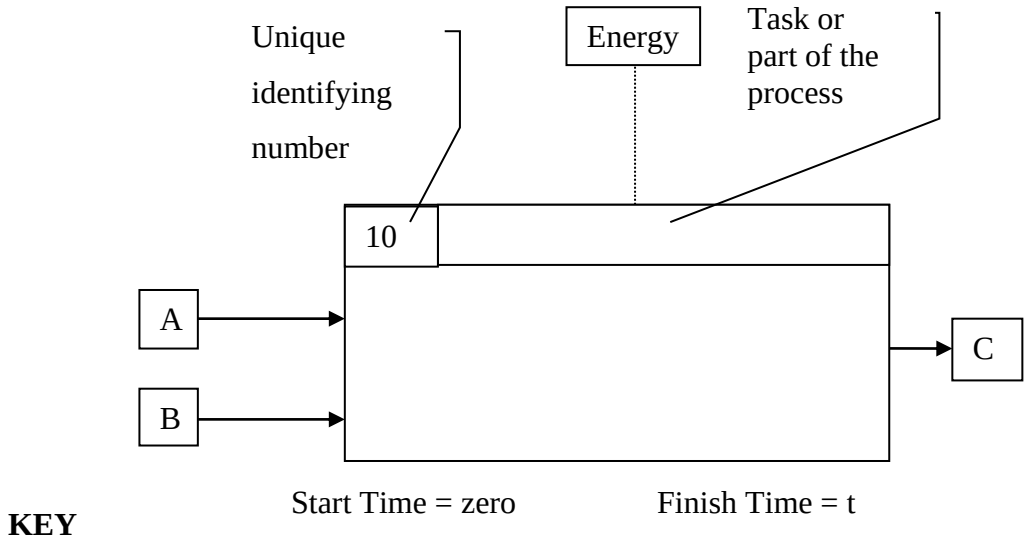


Figure 3-21 PDD Model Pre adaptation (Britest, 2015).

Figure 3-21 shows information captured on the model in a given situation. This can include the state of the process, such as whether wet solids are present, aqueous or solid phases, organics present and also energy used. It informs the user what conditions occur in the vessel during a process, such as a cleaning process.

In order for this tool to be more useful for visualising cleaning in vessels, adaptations can be made, converting the PDD into a Process Definition Cleaning Diagram (PDCD) Figure 3-22.



KEY

A Feed stream chemicals/ type of water used **B** Feed stream detergent used

C Waste Stream (denote waste type) and type of packaging required or used.

Change the vessel box colour to represent the condition, geometry complexity and age of the equipment.

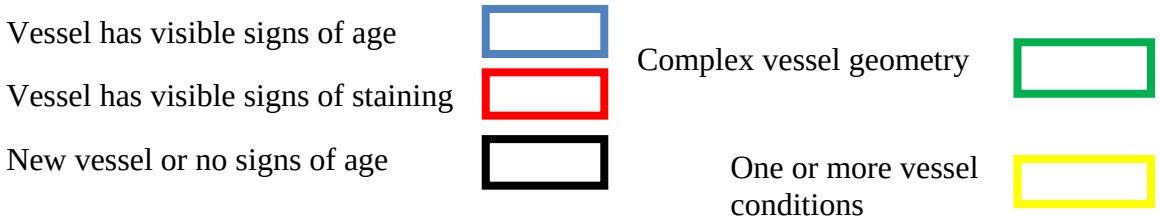


Figure 3-22 An adapted Britest PDD model known as PDCD.

In addition to the changes indicated in figure 3-22 it may also be possible to include details of sampling for cleaning purposes and temperatures in the vessels.

It can be determined that the use of a Rich Picture (RP) could be used at this point to help show specific staining on a vessel in association with the PDCD. This would give a whole vessel overview on those vessels that provide specific cleaning challenges.

In addition to adaptations to the PDD it may be necessary to consider using transformation maps (TM), which are used by Britest members to determine chemical transformations and adaptations taking place in their processes. This is an important tool for use in helping to

optimise processes, but used in the context of cleaning it can help determine potential residues and contaminants left in vessels (shown in the PDCD or RP) which require removal.

Key contaminants not considered during the development of a process can be identified by using a transformation map. Many contaminants are present post reaction therefore identification of these residues or contaminants may help understand how they can be removed by use of solvents or detergents. This part of the analysis is complex and requires understanding chemistry, not just in the cleaning but in the processes themselves. Transformation maps are able to show process transformations in green, and negative non value added transformations of side reactions in red.

Once this information has been processed, the database can be populated with any relevant information and the true cost of cleaning can be determined, which leads to the identification of cleaning process improvements. ZEAL database adaptation has been achieved by determining cleaning information using modified Britest tools as described. However, it is considered that there is a gap in understanding the fundamental science behind cleaning, which is necessary to fulfil the requirements of Britest industrial members and the remit of this research.

3.8 Chapter 3 Summary

Chapter 3 has provided further evidence of the need for a greater understanding of plant cleaning. It is now considered that most cleaning related information is contained in industry and due to reasons relating to confidentiality it is not often shared outside of companies. It is important to this research project that information relating to cleaning and the understanding of cleaning was gained from both the survey and the survey members.

This was recognised by interpreting the survey results which indicated -

- The understanding of contaminants in process plant was not fully understood but a majority of companies considered the contamiant to be chemical.
- Cleaning protocols were designed with the consideration of a number of factors such as contact time, removal of contaminants and volume of cleaning agent used.
- Process plants generally consist of the same components but the size and complexity can vary. Each process plant can be said to be unique.

- Process plants are cleaned according to the type of plant, the product which has been in the equipment and the product proceeding it, the organisation carrying out the cleaning, and special requirements. Special requirements may include disassembly or targeted cleaning in specific areas.
- Most companies surveyed suggested that their cleaning was sometimes effective, this means cleaning would have to be repeated until cleaning was carried out to the desired level.
- In order to clean equipment companies use a number of cleaning agents which include organic solvents, aqueous detergents, mineral acid or alkali and water.
- The choice of solvent was governed by selection in the laboratory using a non scientific based solvency test after the process was developed and the API was made. The choice of cleaning agent was also made during plant commissioning, which is generally after the process has been transferred into manufacturing. This is not ideal, as if the chosen cleaning agent is not able to clean the equipment, a lot of time and resources could be used to try and clean the equipement at this stage. It is important to prevent this.

In addition, site visits have determined some company specific cleaning challenges which have been useful in considering the direction of the next phase of this research. If plant cleaning is to be more effective it must be considered at an earlier stage of the process than the survey data suggests and using WPU to carry this out. It is also important to consider understanding the fundamental science behind cleaning rather than the solubility profile alone, which does not always make a cleaning agent effective.

Chapter 3, section 3.4 answered research questions RQ 3, RQ 4 and RQ 5 based on the results of the cleaning survey and member site visits.

Britest tools have shown that in a theoretical sense they can potentially be powerful in describing processes, confirming what is known about processes and what is not known. A gap in the Britest tool and methodologies has been identified which is a tool to help understand the fundamental science behind cleaning. This will be considered in the following chapter (Chapter 4).

Identifying cleaning metrics are important and in this chapter cleaning metrics have been considered with regard to the ZEAL project and in terms of the cost of pharmaceutical drugs

produced and solvent costs. It is considered that this is possible, but that the information required in order to operate a tool such as the ZEAL database is difficult to obtain.

It is therefore considered that any tool developed for improving cleaning must consider the following points. It must incorporate the fundamental scientific understanding of either solvent or product (including APIs, API intermediates and side products). The tool must be developed for use with the existing Britest tools which as previously discussed, can provide valuable information of chemicals in the process and reactions taking place, and identify potential cleaning challenges. In addition it is considered that the tool must be easy to understand and used earlier in the manufacturing process, before a product is manufactured at small scale in a laboratory, or large scale in a manufacturing plant. This would give anyone using it a significant advantage. This is because potential cleaning challenges may be identified earlier in the process design. At this point they can be either eliminated or reduced by changing chemicals in a process for different ones and avoiding the production of side reaction compounds or intermediates which may be hard to clean from equipment.

3.9 Conclusions

Chapter 3 indicates that there is a lot of knowledge around process plant cleaning, but the depth of knowledge is not enough. There is a lack of fundamental scientific understanding around process plant cleaning. If the gap in fundamental scientific understanding was addressed it would lead to improvements in choices of cleaning methods which would save time and resources. This chapter has shown that companies may not always be aware of the challenges associated with cleaning until the processing plant is commissioned. This is late to consider the choice of cleaning agents. Even if the cleaning agent is chosen before this stage there is no real scientific understanding behind the choice of cleaning agent. This is because an agent is chosen based on a solubility test in a test tube, which does not reflect real process equipment, or anticipate the challenges associated with cleaning complex equipment. Site visits to Britest members has indicated process plant equipment is complex and can be very challenging to clean. This would indicate that it is considered necessary to consider plant cleaning at a very early stage in the manufacturing process. This is in keeping with the philosophy of understanding WPD and in particular WPU. It is considered that a fundamental scientific understanding of the science behind cleaning needs to be carried out. The development of a tool for these purposes will be considered in Chapter 4, which discusses the choice of methodologies for the development of any tool or methodology to begin to address the challenge of understanding the fundamental science behind plant cleaning.

Chapter 4. Materials and Methods

4.1 Introduction

The previous chapters have discussed the answers to several of the research questions raised at the beginning of this thesis, RQ3 to RQ7 (Section 1.3). It is recognised that there is a need for more fundamental understanding of the science behind cleaning with solvents and other cleaning agents. This chapter describes the methodology used in order to develop a Britest tool, which can be used by industrialists to begin to understand the fundamental science behind cleaning.

While research has been carried out to begin to understand cleaning mechanisms, for example, adhesion and cohesion (Fryer et al 2009), and the mechanics of cleaning such as the most effective cleaning flow rate (Fryer et al 2009), no-one has yet considered trying to understand cleaning using knowledge of the basic chemistry of molecules, and which solvents would be considered the most effective to remove residues from a surface. It seems logical to try and understand cleaning by looking at this aspect, as different chemicals have different molecular structures and different physicochemical properties. Therefore, by understanding the make-up of chemicals such as API's, the physicochemical properties of the structure can be understood. Although this is commonplace in industry when considering how to manufacture products such as API's, this has not been applied to cleaning. In order to begin to understand this, this chapter aims to answer the research question RQ2 - what is meant by the term 'fundamental science behind cleaning', in relation to process plant cleaning?

In order to answer this question this chapter investigates the chemical structures (the relative arrangement of the atoms) of a series of API's, their composition (the various atoms making up the molecule), and physicochemical properties. This information was used to create databases. The fundamental science in relation to process plant cleaning must lie within the fundamental molecular information and the physicochemical characteristics. This will be investigated in order to answer RQ2. Therefore this chapter will initially discuss the formation of databases used in this research, and the methodology used to answer the main research question RQ1 - what would be the best way to increase the fundamental understanding of the science behind cleaning linked to solvent and cleaning agent use? This chapter discusses data recognition and the acquisition of data, (section 4.2), the construction of databases containing

the data, and the data pre-treatment (section 4.3). Section 4.4 discusses the choice of methodology and gives a literature review for multivariate analysis, hierarchical clustering and PCA. The initial methodology used for this research is discussed in section 4.5. Section 4.6 discusses the use of Principal Component Analysis (PCA) to gain further understanding and answers to the research questions which are the main aim of this thesis (Section 1.2 and 1.3). This chapter will be summarised in section 4.7 and conclusions are made in section 4.8.

4.2 Data Recognition and Acquisition

4.2.1 Recognition of data

As the aim of the research is to develop a tool to help understand the fundamental science behind cleaning, fundamental scientific data was required in order to do this. Fundamental data of API's concerns the molecular structure and recognisable structural features. This can also mean the physicochemical characteristics of the molecule. In order to clean equipment according to the proposed method as discussed in Chapter 3, it is necessary to know what chemicals, including products and API's, the equipment was in contact with. The most direct way of finding out about the molecular and physicochemical information relating to Britest members API's was to ask Britest members. Obtaining this information from Britest members proved challenging due to concerns around process and product confidentiality. In order to overcome this barrier the research databases were created using API data, which is publically available on Britest member's websites. These API are largely generic molecules of which there is a lot of information available in the open literature. This not only overcame the challenge but it meant that a lack of any data on intermediate chemicals and side reactions from processes reduced the complexity and the size of the data sets for the initial analysis and methodology selection.

A list of 75 API's was identified from Britest member's websites in the public domain. The list was expanded to include several API's from non Britest members which were included in order to increase the amount of data. Once the list of API's had been determined it was necessary to find molecular, structural and physicochemical information for each one. In total 81 API's were selected.

4.3 Database construction and Data pre-treatment

Before database construction and data pre-treatment is discussed it is necessary to discuss the importance of pre-processing of data. Pre-processing of data is a very important first step in

analysis. Any data may be used, as it is as Wold says, 'in statistics it is customary to put all data into a matrix and analyse the lot....all data reflect legitimate phenomena'. Wold (1987) also makes it clear that outliers can severely influence PCA and efforts should be made to remove them from the data set. This is somewhat contradictory and consideration of both points will be carried out for any analysis of the datasets used in this analysis. Other types of data pre-screening which can be utilised are transformation or data expressed as percentages. Transformation of data makes the data more symmetrically distributed, this kind of pre-treatment is common with chromatography data (Wold, 1987). The data may also be auto scaled, means centred, or normalised. The best method to use for specific data can only be determined by trying a few pre-processing methods and selecting the best (Zitko, 1994). Unfortunately, there is still a lot to understand about pre-processing methodologies and how they can be used most effectively (Praveena et al 2012).

Molecular, structural and physicochemical information for each API identified for this research was required in order to create databases of information. It was recognised that this would generate a lot of data, as there are a large number of variables associated with API's. Therefore, early in the research it was established that three databases were needed. It was determined that the best method to use for pre-processing the data was normalisation. This was because databases 1 and 2 needed to be combined. The variables all had different units and without using this technique analysis would be difficult. Normalisation gave each of the variables an equal weight. This is further discussed in section 4.3.5. Descriptive statistics are provided in Appendix IV for database 1 and database 2 in order to help characterise the databases.

4.3.1 Database 1: Chemical functional groups

The first database was constructed using information on chemical functional groups for the API's identified. This was carried out by visual inspection. An example of how the chemical group information was obtained is given for the drug Aspirin (figure 4.1). Initially the structure of the API was identified by entering the generic name of the API or the company specific name for the API into software. The software used for this purpose was ChemSpider (Chemspider, 2015) or ChemDraw (2015). This gave the structure of the API providing information used for the research. This data matrix was composed of variables associated with the chemical functional groups of the API. This includes information such as the type and number of amine groups present, the type and number of carbonyl groups present, or any structural features such as an organic framework. A list of variables used in database 1 is given in Appendix IV. The database is also shown in Appendix IV.

Figure 4.1 Chemical functional groups in Acetylsalicylic acid (Aspirin) Molecular Formula $C_9 H_8 O_4$ (Pubchem, 2015).

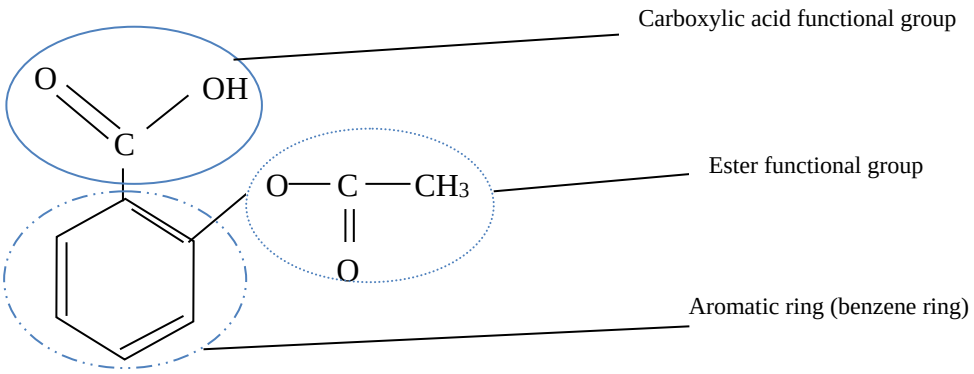


Figure 4.1 Chemical functional groups found in Aspirin.

Each of the chemical functional groups identified as indicated in figure 4.1 would be recorded in an appropriate column in an Excel spreadsheet. This methodology resulted in the generation of data on chemical functional groups. The variables were tabulated and stored in an Excel spreadsheet (Excel, 2007). Figure 4.1 shows that Aspirin would be recorded as containing one carboxylic acid group, one Ester functional group and one aromatic functional group (benzene ring). Tabulation of the data for each API recorded multiple instances of chemical functional groups in some cases. The creation of this database allowed analysis of the compositions of each API.

4.3.2 Database 2: Physicochemical properties

Construction of the second database also required the use of the software ChemDraw and Chemspider, as it required information on physicochemical properties (variables) of the same 75 API's. Physicochemical information for the purposes of this thesis included information specific to properties such as melting point of the API, information on Henry's Law, Gibbs Law, and many other characteristics that were identified from the indicated software.
The information was collated and entered into a database in Excel to enable analysis.
A list of variables used in database 2 is given in Appendix III. The database is also given in Appendix IV.

4.3.3 Database Three

Database three was constructed from database one and database two. The data contained in both databases was combined into one excel spreadsheet in order to make it easier for analysis. This database was therefore a complete set of all of the data collected.

4.3.4 Database Information

It is important to state that not all of the required information concerning chemical functional groups, structure or physicochemical characteristics was available for the entire set of identified API's. (This led to challenges concerning missing data during analysis which will be discussed later in Section 4.4.1).

It is also important to state that it was not known at this point if all of the information collected in both databases was of relevance to this research.

4.3.5 Data Pre-Treatment

Due to the nature of the data, databases one and two required pre-screening prior to analysis. It was important to treat both databases in the same manner in order to allow easy merging of the databases for analysis of all data variables in database three.

Pre-screening involved assessment of the data and required the removal of some pharmaceutical product information, which contained limited data from the product list to ensure that data fields were as complete as possible. As previously stated it was not possible to obtain data for all identified API for this research. For the initial analysis the databases were used with data gaps as indicated in section 4.4.1. It was therefore necessary to omit some API's from the analysis due to a lack of data availability after this initial research. This reduced the number of API's used in this analysis. This list of API's is given in Appendix II.

The data sets were also normalised to obtain values between 0.0 and 1, using the following calculation (equation 4.1).

(V - mean of V)/s

Equation 4.1 Normalisation calculation

Equation 4.1 can be explained as follows. V is the variable dataset. V is divided by the mean value of every variable data set. The result is then divided by s which signifies the standard deviation of every variable.

It was important to introduce normalisation because of multiple measurement parameters or scales involved in the physicochemical database. It was also important as the first and second databases were to be combined to create the third database. Other pre-treatment of data was not carried out in this research although the researcher is aware of other techniques (discussed in section 4.3).

Once pre-treatment was carried out for each variable database, analysis could then begin. The next step was to determine which methodologies to use to analyse the data. This is discussed in section 4.4.

4.4 Choice of Methodology

This section discusses the reasoning for using multivariate analysis. The data collected for this research project was complex due to the multiple variables, and any methodology chosen to analyse the data would need to be multivariate, to allow visualisation of the interactions of the variables in the data set. However, there are several types of multivariate analysis. Hanley, (1983) gave an overview of multivariate analysis and described it as "a collection of statistical techniques for dealing with several data items in a single analysis". This concerned data defined as 3 variables and above analysed at the same time. Hanley, (1983) also described the method of analysis chosen to be dependent on "whether one is interested in interrelationships or in comparisons, and on whether variables are qualitative or quantitative." The data collected for this research thesis was quantitative and it was important to consider the interrelationships. In order to select the best methodology it was necessary to carry out research into the uses of each method by carrying out a literature review (section 4.4.1) followed by trying different methods with the data.

The software package available for use for the analysis was Minitab, as licenses were provided for this it seemed logical to use this software over other packages such as R or SPSS. Minitab software has several options available for analysing data. Analysis of data structure by covariance can be carried out using Principal Component Analysis (PCA) or Factor Analysis. Prior to using a technique to examine the data structure it was considered important to look at the data using a grouping method or cluster analysis to see if any immediate patterns or clustering was present. This was deemed a "quick and dirty analysis". The types of methods available for this were the grouping observations. There were also several options available including cluster observations, cluster variables and cluster K-means. This was carried out to see if anyone had carried out similar research in this field and also to determine what other researchers were using to using these methodologies for. This is discussed in the next section 4.4.1.

4.4.1 Literature Review of Methodologies

This section will initially discuss some of the theory associated with Multivariate analysis. Multivariate analysis has been described as a "codification of techniques of analysis, regarded

as attractive paths rather than straightjackets, which offer the scientist valuable directions to try" (Bishop et al, 1976). Carrying out an initial literature review on multivariate methodology it quickly became clear that these techniques are carried out in a number of scientific and social science research areas for multiple purposes. This observation fits in with Bishops definition that analytical paths have been chosen by many people for many purposes. Considering the data available for the research it was important to ask what the outcome of the research was. For this thesis it was important to discover if any patterns or links could be found in the data which would indicate how chemicals could be cleaned from process equipment. Any links and patterns may indicate ease of cleaning or difficulty in cleaning chemicals. It may help discover new cleaning methodologies for chemicals that are difficult to remove from vessels. Therefore the interaction between the data was important and therefore all variables in the analysis were treated equally. This meant that any technique used needed to take this into account. Techniques which can do this are PCA, Factor Analysis and Cluster Analysis. This narrowed down the techniques used in this thesis to these three techniques. Where cluster analysis groups objects based on a measure of proximity and classification, PCA and Factor analysis do not classify data but do reduce its dimensionality. Factor analysis and PCA are similar techniques but the aim of the thesis was to provide a useful technique to identify cleaning agents for the Pharmaceutical industry. It was important to use PCA as the method of choice because it reduces the number of variables to those which give the most variation in the data set. For this reason the literature review will concentrate on the use of Cluster Analysis (section 4.4.2) and PCA (section 4.4.3).

4.4.2 Literature Review of Hierarchical Cluster Analysis

This section discusses the theory behind cluster analysis and will also discuss some uses of the technique. Cluster analysis, initially mentioned in section 4.4, is further described here. Clustering methods are used in exploratory data mining and are also a common technique in statistical data analysis. There are many types of clustering methods including the hierarchical methods, portioning relocation methods, grid based methods and density based partitioning methods (Kogan, 2006). This thesis will use hierarchical clustering methods (also called connectivity based clustering), so this literature review will focus on these techniques. There are two types of Hierarchical clustering methods. The first is agglomorative and the second is divisive. Agglomorative clustering describes "bottom up" clustering where at the start of the analysis each variable belongs in its own cluster. During analysis clusters merge until one cluster remains for example using the Sequential Agglomerative Hierarchical Non-overlapping (SAHN) technique by Sneath and Sokal (1973). The second method is described

as “top down” where one cluster is present at the beginning of the analysis and then it is divided successively until each item is in its own cluster. Clustering for each method continues forming appropriate sub-clusters until a stopping criterion is achieved. Stopping criterion will be discussed later in this thesis.

There are advantages and disadvantages to using hierarchical clustering. Advantages include that the technique can be used for any attribute type, the techniques shows a degree of flexibility regarding the level of granularity and its similarity or distance in any form can be handled with ease. There are several disadvantages to using this method which must be taken into consideration. These include the difficulties in choosing the correct stopping criteria and the fact that most hierarchical algorithms do not revisit (intermediate clusters) once they are constructed (Kogan, 2006). In addition, divisive clustering is thought to be more sophisticated and provides more robust clustering (Izenman, 2013).

The results for both types of hierarchical clustering of data are best shown in a dendogram. The choice of clusters is made by algorithms which produces a hierarchy. Hierarchical clustering deals with N x N matrix of distances which can be similarity or dissimilarity between data points. One of the most difficult aspects of using clustering techniques is cutting or partitioning the data in the dendrogram at a certain height to give a partition of the data. This is also known as stopping criteria. This is carried out by calculating the distance between data points. The most commonly used method to measure distance is the Euclidean distance metric, which measures the geometric distance in the multidimensional space. This was used for this research thesis, as the data variables in the database chosen were all in the same physical units. Stopping methods for optimising clustering is a fundamental problem and it is challenging. Decision rules do exist and have been provided by Milligan and Cooper (for agglomerative clustering) to determine the appropriate level of the dendrogram. Principal Direction Divisive Partitioning (PDDP), a dynamic threshold based method, aims to stop the partitioning when the centroid scatter value exceeds the maximum cluster scatter value at any particular point (Jung et al, 2002).

Choice of linkage is the other decision a researcher needs to make when using hierarchical clustering. This shows patterns and gives structure to the data. The choice of a linkage method determines how the variables are shown. On a dendrogram the height of clusters indicates the similarity or dissimilarity. Similar variables are shown at low heights, while dissimilar variables are shown by a difference in height. Clustering can be difficult as there is no right answer. Clustering will be discussed further later in this section.

There are multiple possibilities for linkage the most common methods include complete linkage (furthest neighbour) or single linkage (nearest neighbour), the average link method, and Ward's method (1963). The method chosen for this research was single linkage. This looked at linking data when any two variables in two clusters were closer together. The problem with this linkage method is that it can chain data together, called "chaining", where a sequence of close observations in different groups can cause the groups to merge early.

There are several methods for cutting or splitting data in an agglometric dendrogram. The oldest method is by Williams and Lambert (1959). In this method objects are split based on the values of only one variable. In the Macnaughton –Smith et al. (1964) method, a split is decided by taking the most distant object from the cluster for a new cluster. Other objects are then aggregated to the new cluster if they are closer to the new subset than the cluster they are in. This idea is similar to Huberts (1973) method, which takes a pair of objects which are most dissimilar to the cluster for the new cluster. The new clusters are then built according to distances between the original cluster and the new pair. In 1991 Roux exploited this idea and considered clustering generated by all pairs of objects, creating a priori like criterion.

It is important to treat cluster analysis with caution, as different decisions concerning similarities in groups can give different dendrograms. Sometimes a hierarchical structure is imposed on the data even if it is not appropriate. Sorlie et al (2003) used the method to examine patterns of gene expression for clinical classification of tumours. The clustering lead to new theories which found that some breast tumour subtypes represent distinct biological entities, but in the profiling the data observed in clinical samples disproved these theories.

Clustering is a popular method for analysing data in many fields for example in medicine (Boly, 2012, Leite, 2015) and in astrophysics which is discussed below.

Cluster analysis can be carried out by cluster observations, cluster K-means, or clustering variables. Clustering observations are a useful technique when there is some information available about potential clusters. This methodology is popular in several fields of research including in space physics and geophysics, where researchers have used the technique to analyse Electromagnetic ion cyclotron (EMIC) waves in plasmaspheric plumes. This determined that cold plasma density was not a good predictor of EMIC occurrence inside plumes (Usanova, M.E 2013). Researchers have also used this technique to determine density irregularity in the plasmasphere boundary layer (Decreau et al, 2005). The methodology has been used to determine star clusters in galaxies by several researchers including Hattori et al,

(1997) and Joyce, M et al (2015). Cluster observation has also been used in many other fields including research into water quality (Singh et al, 2004), and Vialle, C et al, (2011).

Cluster K (Number of K clusters means) methodology has been used in several different fields by researchers. It appears to be a common methodology in public health, where researches have used it, for example, to determine the dietary patterns of middle aged Irish men and women (Villegas, 2004) . Cluster K means has also been used by sports scientists to determine the motivational orientations and imagery use in goal profiling (Cummings, 2002).

Clustering variables is a method which is used when there is no obvious relationship or grouping in the data. This method is frequently used by researchers in many fields, including earth sciences, to determine the physical and chemical variables in soil for example, by Arslan, (2013) and Irigoin (2016).

There was limited available literature on using these techniques for pharmaceutical cleaning purposes, or for trying to group or cluster chemicals based on chemical functional groups or physicochemical properties. Two research groups used multivariable techniques for similar purposes, to determine clusters in data relating to pharmaceutical solvents (Xu, 2007), and analysis of collections of chemical compounds to identify potential lead drug candidates (Stanton, 1999).

4.4.3 Literature Review Principal Component Analysis

This section will focus on the theory behind principal component analysis before discussing some of the uses of PCA. PCA was initially developed by Pearson who described it *"as finding lines and planes of closest fit to systems of points in space"* (1901). This technique was further developed by other researchers, notably Wold (1987) and Hotelling (1933).
PCA is a multivariate analysis mathematical methodology for reducing a large database of interrelated information or variables to a reduced number of principal components. The aim of the analysis is to show or explain the maximum amount of variance within the data set with the least number of principal components. A principal component is a new latent variable and all principal components are linearly uncorrelated to others. Principal components are ordered during the analysis so that the first few principal components retain the most variation which is present in the original variables (Jolliffe, 2002). The variance in the original data can be expressed as linear combinations of the principal components, i.e. -

$$X=P*T$$

Where X is the data matrix and P and T are two smaller data matrix which capture the variability or the essential data in X. "*Plotting the columns of T gives a picture of the dominant 'object patterns' of X and, analogously, plotting the rows of P shows the complementary 'variable patterns'"* (Wold, 1987).

There is more than one type of PCA, including Common PCA, Functional PCA, Multiway PCA and Rotated PCA, and details of how these techniques have been used by researchers is provided below with their reasoning for use of the technique where appropriate.

Principal Component Analysis (PCA) has been used in a number of diverse fields to analyse multiple variables, find patterns in data and use the information to find the best solution for storage (such as food or drink) or interpret variables. One of the most challenging aspects of PCA is understanding which components to retain. Webster (2001) says judgement must be made when choosing which components to retain and many tests proposed for this purpose are at best guides. Literature has shown that different fields favour different types of PCA and different methods to choose components to retain. Examples of research using PCA to analyse data, show different techniques and the breadth of its use across disciplines are described below.

PCA has been widely used in social sciences to understand and analyse water chemistry (Dong, 2007), to determine mineral composition in Cigua (Oliveira et al, 2014), and to determine stream and water chemistry conditions in waste water (Wallace and Champagne 2013).

PCA has been used extensively in science, for example, Maere et al (2012) used the technique along with fuzzy clustering to determine bioreactor fouling behaviour. They used PCA (common PCA and types of functional PCA (expert PCA and B-splines PCA) to analyse transmembrane pressure data. It was possible to use functional PCA as the data set was known well. This technique required data conversion into a set of function parameters (separately estimated for each data series prior to PCA). This resulted in scores which captured the most variance in the data. Using this technique the results of the PCA analysis shift from the raw data to parameters of the functions. This technique has several advantages over common PCA which are that the estimated functions are able to express expert knowledge. This makes the data easier to interpret. The choice for one particular function means that analysis can focus on variations of interest, and also the number of PCA parameters using this model is generally lower than the number of variables in each series (than in common PCA). (Maere et al, 2012). This research allowed them to choose the best

method to use for this type of data. The preference was for expert PCA, as it handled outliers and noise better than the common PCA and it is less complex than the B splines method.

PCA has also been used in food science, for example to evaluate the aroma quality of Chinese traditional soy paste during storage (Peng et al, 2014). Using PCA analysis the researchers were able to determine 15 volatile components in the samples and place them into three groups (overall odour types in storage periods which were floral roasting and pungent) based on the distribution on the factor loading plot. PCA has been used to differentiate gelatine sources based on polypeptide molecular weights (determined by sodium dodecyl sulphate (SDS) -PAGE). Analysis by PCA showed that using the molecular weights of gelatine it was possible to determine 5% porcine gelatine in bovine gelatine. This technique is useful for determining the purity of gelatine in a product.

PCA is used in the chemical and manufacturing fields for many applications, one of which is described below. Nomikos and MacGregor (1994) used a non-linear PCA technique (multiway principal component analysis) to track batches of product in manufacturing. This is important as it ensures safe operation and to ensure that high quality products are produced. Nomikos and MacGregor (1994) did this using historical data of successful batches and compressing the data onto a low-dimensional space that summarizes both the variables and their time histories. A new batch of chemical could then be monitored by comparing its progress against the normal previous successful batches (Dong and McAvoy (1996)). Dong and McAvoy subsequently used a different method, Non-linear PCA, to successfully track batches of product in a manufacturing environment. Non-linear and linear PCA are the same apart from the fact that the non-linear approach summarizes the data with a smooth curve that is determined by nonlinear relationships among all the variables. Most batch data is non-linear and using a non-linear methodology was shown to have advantages over the multiway PCA. This is because it is thought to compresses data more efficiently. The multiway PCA is considered cumbersome if more than three components are needed to describe the data, as there are too many plots to analyse (Dong and McAvoy, 1996). In addition, using the multiway method may mean that the results are inadequate as minor components may be discarded. These minor components may contain important information. PCA has been used in other fields but research linked to this thesis topic has been difficult to find. Some analysis of data has been carried out in chemistry, for example Malinowski used PCA to analyse proton shift of methanes in a variety of solvents with (Trimethylsilyl groups) TMS (1970).

4.4.4 Cluster Analysis

Selecting clusters in a dendrogram, PCA or other analysis techniques has been approached by researchers in many different ways. There is no defined optimal method and many consider the results of clustering as misleading (Gordon 1996). Many researchers have created new ways to cluster data as they try and look for more efficient methodologies, and importantly, validating the results of their analysis. These include stopping rules which define when to stop clustering data (Howe, 1979, Legendre et al 1985), or augmentation of a single class, the simultaneous test procedure developed by Gabriel and Sokal (1969). This provides a bound for the probability of incorrectly subdividing any class which is specified as homogeneous by a statistical model (Gordon, 1996).

How researchers select the number of groups in clusters can be carried out by a number of algorithms or methods. Many of the methods are developed in specific fields of research where difficulties present themselves when identifying clusters. For example, Guidi et al, (2009) proposed the use of random simulation test (RST) proposed by Ibanez (1973) to identify meaningful principal components. The RST takes into account if the data set contains statistical outliers and if they are present it isolates them instead of clustering them. This method was later used by Nicolls et al (2010) for the determination of the optimal number of clusters to be extracted from a classification used in ecological studies.

Cluster analysis is a popular technique in operations management to determine manufacturing strategy taxonomy by Miller and Roth (1994), and Wathan (1995). Since the early 1990s the method has been used many times. Less traditional methods are being used in this field as more emergent techniques become available. These include p-median clustering, model-based clustering, neural network clustering, overlapping and fuzzy clustering and network clustering. It is considered by some that these emergent methods are not a replacement for the traditional clustering methods, but are suitable alternatives for some applications (Brusco, 2012).

In this section the uses of Hierarchical clustering, PCA and choice of clusters have been examined. The following two sections will discuss the initial method development using hierarchical clustering (section 4.5) and PCA (section 4.6).

4.5 Initial Method Development - Hierarchical Clustering

4.5.1 Initial Method development - Multivariate analysis

The initial methodology development was carried out using the information of the physicochemical properties (data in database 2). The most important criteria for choosing a methodology to analyse the information was the multivariate nature of the data. This meant that the initial methodology considered for analysing the data for groupings and clustering effects was multivariate analysis. The first analysis technique chosen was a hierarchical clustering system with the aim of clustering the variables into groups. This was chosen over other methods because potential groupings were unknown and no information on how the data may be grouped existed. The aim of clustering is to find an optimal grouping where the variables in each group are similar, but the clusters are also different from each other. The resulting groupings in the data are ones which the researcher can see are sensible and can make sense of (Rencher, 2002). Hierarchical clustering in this case is carried out where a number of variables start out at the beginning of the analysis as discrete clusters. As the analysis progresses, the number of discrete clusters decreases as similarities are found between the data. This resulted in a hierarchy of clustering in the data, where cuts in the data can be made according to the relationships found.

Using this methodology in Minitab (version 16), it is important to choose the correct linkage option. The linkage type option chosen for this technique was single linkage (nearest neighbour). This was because it identified groups which were spatially close in the data. The resultant clustering is visualised in a dendrogram generated in Minitab with database two, which shows the similarity and patterns within the data (Chapter 5, section 5.2). The main challenge associated with using this methodology was the amount of missing data. This was due to the nature of the data itself, and the fact that for some of the pharmaceuticals initially listed, it was not possible to obtain or generate data needed for the analysis. Some of the variables, which were included in the research data, were not calculable for every API.

Database 1 initially gave 64 variables, which are listed in appendix III. This list was reduced to 57 variables by the exclusion of the structural and molecular features listed - sulfonated molecules, aldehydes, anhydrides, expoxides, nitriles and thiol.

Database 1 initially contained 81 API’s. This was reduced to 73 during the analysis, due to the limitations of available data.

The data in database 1 was analysed with PCA (as was the data in database two). The use of PCA as a methodology will be described in section 4.6.

4.6 Principal Component Analysis

4.6.1 Principal Component Analysis Examination as a methodology

This section builds on the knowledge gained from the previous section 4.4. Due to the number of variables in the datasets, it was important to consider another multivariate analysis technique to analyse the data. The next technique considered for examination of the data was Principal Component Analysis (PCA). PCA was used in this research because it was important to determine the interrelatedness between and within the variables. Due to the complexity of the databases, this type of methodology would reduce the number of variables considered as significant. It would compress the data and filter out some of the noise within it. The main aim of principal component analysis is to give reasons for the amount of variance in a data set with the fewest number of principal components. The principal components can be defined as *"linear combinations of the original variables calculated with the maximum variance criterion. Principal components are centred, uncorrelated, and ordered from the largest to the smallest variance"* (Minitab, 2016). The first principal component is the linear combination of all of the x-variables that comprise the maximum variance amongst all of the data.

4.6.2 Principal Component Analysis of the data

The data sets described in section 4.3 were entered into Minitab software (Version 16) and PCA was carried out. For this analysis, correlation or covariance could be used to measure the strength between two random variables, looking for patterns and linkages in the data set. Covariance was chosen for this analysis because it is a measure of the strength of the correlation, and not the strength of the linearity between the variables. The analysis was initially carried out with database 1 (on normalised data), which contained information on the variables associated with the chemical functional groups, and which were identified in the chosen list of API's. A list of the chemical functional groups and properties chosen for this analysis is given in Appendix IV. Analysis was performed by PCA to help establish links and clustering effects between pharmaceutical products and the chemical functional groups. The results of this analysis are shown and discussed in Chapter 5 of this thesis.

Further analysis was carried out on each of the remaining databases (two and three). The results of both of these analyses are shown and discussed in Chapter 5.

4.7 Chapter Summary

This chapter discussed the origin of the data, which was required for the research. It was important to consider what information would address the challenge of beginning to understand the fundamental science behind cleaning. It was considered that the fundamental data must involve characterisation of the API's that Britest members produce. This included obtaining information on the chemical functional groups and structural features which the API's all consist of. Each structural and chemical make-up is unique to each API. This means that it might be possible to use a multivariable analysis technique allowing features to be clustered and grouped. This could be considered as a methodology to group API's together for analysis with respect to cleaning purposes. In addition this analysis must include the fundamental physicochemical information, as it may also be able to indicate where API's may be clustered together for analysis with respect to cleaning purposes.

Carrying out a literature review on multivariate analysis techniques, and in particular Hierarchical clustering and PCA, has shown that trial and error often leads to the selection of the correct multivariate technique to use to analyse data. It is considered that one of the most important aspects of carrying out analysis using multivariate techniques is to understand how to examine clusters of data on plots produced during analysis. This will be considered when analysing data in this thesis.

4.8 Conclusions

This chapter identified the methodology to use to carry out the research as PCA. This is due to the nature of the data as discussed in section 4.4.

The answer to research question RQ2 (Chapter 1, section 1.3), i.e. What is meant by the term 'fundamental science' in relation to process plant cleaning, has therefore been partially identified for the purposes of this research. Chapter 5 seeks to further identify an answer to this question by focussing on the results obtained from the analysis of the three databases. This will lead to the identification of key variables that can indicate the best methodologies, which can be used to clean process plant equipment post manufacturing specific APIs.

The results of the analysis on each of the databases will be discussed in Chapter 5 and consideration will be given to the construction of a model which can be used to meet the primary aim of this research, to develop an understanding of the fundamental science behind process plant cleaning. The production of a model to increase the knowledge of process cleaning will be a tool to help Britest members understand cleaning. This will be discussed in Chapter 5. Chapter 5 will also indicate where the tool will fit into the Britest tool set that was

identified in Chapter 3 as of considerable use to Britest members seeking to understand the fundamental science behind cleaning.

Database 1 and Database 2: Raw Data

Database 1: Chemical Functional Groups

	A	B	C	D	E	F	G	H	I	J	K	L	M
1		Amine				Alcohol OH					Acid		
2													
3		Primary	Secondary	Tertiary	Aromatic/enamine	Primary	Secondary	Tertiary	Vinyl alcohol	Phenol	Carboxylic	Sulfonated	Other?
4	Advicor (Niacin + Lovastatin)	0	0	0	0	0	0	0	0	0	0	0	0
5	Aluvia (Lopinavir + Ritonavir)	2	0	0	0	0	0	0	0	0	0	0	0
6	Androgel (Testosterone)	0	0	0	0	0	2	0	0	0	0	0	0
7	Atenolol	0	1	0	0	0	1	0	0	0	0	0	0
8	Bambec (Bambuterol)	0	1	0	0	0	1	0	0	0	0	0	0
9	Beclomethasone dipropionate	0	0	0	0	0	1	0	0	0	0	0	0
10	Beclomethasone dipropionate monohydrate	0	0	0	0	0	1	0	0	0	0	0	0
11	Betamethasone acetate	0	0	0	0	0	1	1	0	0	0	0	0
12	Betamethasone disodium phosphate	0	0	0	0	0	1	1	0	0	0	0	0
13	Blopress (candesartan cilexetil)	0	0	0	0	0	0	0	0	0	1	0	0
14	Brufen (Ibuprofen??)	0	0	0	0	0	0	0	0		1	0	0
15	Calcijex (Calcitriol)	0	0	0	0	0	2	1	0	0	0	0	0
16	Carisoprodol	0	0	0	0	0	0	0	0	0	0	0	0
17	Ciclesonide	0	0	0	0	0	1	0	0	0	0	0	0
18	ciclosporin	0	0	0	0	0	1	0	0	0	0	0	0
19	Citanest (Prilocaine)	0	1	0	0	0	0	0	0	0	0	0	0
20	Clarithromycin	0	0	1	0	0	3	1	0	0	0	0	0
21	Clobetasol propionate	0	0	0	0	0	1	0	0	0	0	0	0
22	Conholip	0	0	0	0	0	0	0	0	0	0	0	0
23	Cycloserine	1	0	0	0	0	0	0	0	0	0	0	0
24	Deflox (Terezosin hydrochloride)	0	0	0	1	0	0	0	0	0	0	0	0
25	Dexamethosone dipropionate	0	0	0	0	0	1	0	0	0	0	0	0
26	Doxycycline hyclate	0	0	1	0	0	1	1	3	0	0	0	0
27	Doxycycline monohydrate	0	0	1	0	0	1	1	3	0	0	0	0
28	Epival (Sodium valproate)	0	0	0	0	0	0	0	0	0	1	0	0
29	Eprosartan (Teveten)	0	0	1	0	0	0	0	0	0	2	0	0
30	Fluticasone furaroate	0	0	0	0	0	1	0	0	0	0	0	0
31	Fluticasone propionate	0	0	0	0	0	1	0	0	0	0	0	0
32	Folic Acid	1	0	0	1	0	0	0	0	0	2	0	0
33	Furosemide	0	1	0	0	0	0	0	0	0	1	0	0
34	Gabapentin	1	0	0	0	0	0	0	0	0	0	0	0
35	Gadopentetate dimeglumine	0	2	3	0	2	8	0	0	0	5	0	0
36	Gadopentetate monomeglumine	0	1	3	0	1	4	0	0	0	5	0	0
37	gopten (Trandonapril)	0	1	0	0	0	0	0	0	0	1	0	0
38	Halobetasol	0	0	0	0	0	1	0	0	0	0	0	0
39	HPMPC (Cidofovir)	0	0	0	1	1	0	0	0	0	0	0	0
40	Hytrin	0	0	0	1	0	0	0	0	0	0	0	0

	A	B	C	D	E	F	G	H	I	J	K	L	M
1													
2		Amine				Alcohol OH					Acid		
3		Primary	Secondary	Tertiary	Aromatic/enamine	Primary	Secondary	Tertiary	Vinyl alcohol	Phenol	Carboxylic	Sulfonated	Other?
41	Imdur (Isosorbide mononitrate)	0	0	0	0	0	1	0	0	0	0	0	0
42	Iodixanol	0	0	0	0	4	5	0	0	0	0	0	0
43	Iohexol	0	0	0	0	3	3	0	0	0	0	0	0
44	Iopamidol	0	0	0	0	4	1	0	0	0	0	0	0
45	Isoflurane	0	0	0	0	0	0	0	0	0	0	0	0
46	Isradipine	0	0	0	0	0	0	0	0	0	0	0	0
47	Ivermectin	0	0	0	0	0	2	1	0	0	0	0	0
48	Ketoprofen	0	0	0	0	0	0	0	0	0	1	0	0
49	Klacid (Clarithromycin)	0	0	1	0	0	3	1	0	0	0	0	0
50	Levothyroxine	1	0	0	0	0	0	0	0	1	1	0	0
51	Lupron (Leuproreline)	0	0	0	3	1	0	0	0	1	0	0	0
52	Marcaine (Bupivacaine)	0	1	0	0	0	0	0	0	0	0	0	0
53	Meperidine	0	0	1	0	0	0	0	0	0	0	0	0
54	Meprobamate	0	0	0	0	0	0	0	0	0	0	0	0
55	Methohexital	0	0	0	0	0	0	0	0	0	0	0	0
56	Metolazone	0	1	0	0	0	0	0	0	0	0	0	0
57	Mometasone furoate anhydrous	0	0	0	0	0	1	0	0	0	0	0	0
58	Mometasone furoate monohydrate	0	0	0	0	0	1	0	0	0	0	0	0
59	Nimbex (Cisatracurium besilate)	0	0	0	0	0	0	0	0	0	0	0	0
60	Nizatidine	0	0	1	2	0	0	0	0	0	0	0	0
61	Olanzapine	0	0	2	0	0	0	0	0	0	0	0	0
62	Oxis (Formoterol)	0	1	0	0	0	1	0	0	1	0	0	0
63	Paricalcitol (Zemplar)	0	0	0	0	0	2	1	0	0	0	0	0
64	Plendil (Felodipine)	0	0	0	1	0	0	0	0	0	0	0	0
65	Progesterone	0	0	0	0	0	0	0	0	0	0	0	0
66	Quinapril	0	0	0	0	0	0	0	0	0	1	0	0
67	Ranitidine	0	0	1	2	0	0	0	0	0	0	0	0
68	Roxithromycin	0	0	1	0	0	3	2	0	0	0	0	0
69	Salmeterol xinafoate	0	1	0	0	1	1	0	0	2	1	0	0
70	Sevelamer	2	2	0	0	0	0	1	0	0	0	0	0
71	Severane	0	0	0	0	0	0	0	0	0	0	0	0
72	Sumatriptan base	0	1	1	0	0	0	0	0	0	0	0	0
73	Tamsulosin	0	1	0	0	0	0	0	0	0	0	0	0
74	Venlafaxine	0	0	1	0	0	0	0	0	0	0	0	0
75	Warfarin	0	0	0	0	0	0	0	1	0	0	0	0

	A	N	O	P	Q	R	S	T	U	V	W	X	Y	Z	AA
1															
2		Carbonyl									Other N groups				
3		Ketone	Aldehyde	Enone	Ester	1o amide	2o amide	3o amide	Anhydride	Epoxide	Thioester	Oxime	Oxazolidinone	Urea	Guanidine
4	Advicor (Niacin + Lovastatin)	3	0	0	0	0	0	0	0	0	0	0	0	0	0
5	Aluvia (Lopinavir + Ritonavir)	0	0	0	2	4	0	0	0	0	0	0	0	0	0
6	Androgel (Testosterone)	0	0	1	0	0	0	0	0	0	0	0	0	0	0
7	Atenolol	0	0	0	0	1	0	0	0	0	0	0	0	0	0
8	Bambec (Bambuterol)	0	0	0	0	0	0	0	0	0	0	0	0	0	0
9	Beclomethasone dipropionate	2	0	0	2	0	0	0	0	0	0	0	0	0	0
10	Beclomethasone dipropionate monohydrate	2	0	0	2	0	0	0	0	0	0	0	0	0	0
11	Betamethasone acetate	2	0	0	1	0	0	0	0	0	0	0	0	0	0
12	Betamethasone disodium phosphate	2	0	0	0	0	0	0	0	0	0	0	0	0	0
13	Blopress (candesartan cilexetil)	0	0	0	0	0	0	0	0	0	0	0	0	0	0
14	Brofen (Ibuprofen??)	0	0	0	0	0	0	0	0	0	0	0	0	0	0
15	Calcijex (Calcitriol)	0	0	0	0	0	0	0	0	0	0	0	0	0	0
16	Carisoprodol	0	0	0	0	0	0	0	0	0	0	0	0	0	0
17	Ciclesonide	1	0	0	1	0	0	0	0	0	0	0	0	0	0
18	ciclosporin	0	0	0	0	0	4	7	0	0	0	0	0	0	0
19	Citanest (Prilocaine)	0	0	0	0	0	1	0	0	0	0	0	0	0	0
20	Clarithromycin	1	0	0	1	0	0	0	0	0	0	0	0	0	0
21	Clobetasol propionate	2	0	0	1	0	0	0	0	0	0	0	0	0	0
22	Conholip	2	0	0	1	0	0	0	0	0	0	0	0	0	0
23	Cycloserine	1	0	0	0	0	0	0	0	0	0	0	1	0	0
24	Deflox (Terezosin hydrochloride)	0	0	0	0	0	0	1	0	0	0	0	0	0	1
25	Dexamethosone dipropionate	2	0	0	2	0	0	0	0	0	0	0	0	0	0
26	Doxycycline hyclate	1	0	0	0	1	0	0	0	0	0	0	0	0	0
27	Doxycycline monohydrate	1	0	0	0	1	0	0	0	0	0	0	0	0	0
28	Epival (Sodium valproate)	0	0	0	0	0	0	0	0	0	0	0	0	0	0
29	Eprosartan (Teveten)	0	0	0	0	0	0	0	0	0	1	0	0	0	0
30	Fluticasone furaroate	1	0	0	1	0	0	0	0	0	1	0	0	0	0
31	Fluticasone propionate	1	0	0	1	0	0	0	0	0	1	0	0	0	0
32	Folic Acid	0	0	0	0	0	2	0	0	0	0	0	0	0	0
33	Furosemide	0	0	0	0	0	0	0	0	0	0	0	0	0	0
34	Gabapentin	0	0	0	1	0	0	0	0	0	0	0	0	0	0
35	Gadopentetate dimeglumine	0	0	0	0	0	0	0	0	0	0		0	0	0
36	Gadopentetate monomeglumine	0	0	0	0	0	0	0	0	0	0	0	0	0	0
37	gopten (Trandonapril)	0			1			1							
38	Halobetasol	2	0	0	1	0	0	0	0	0	0	0	0	0	0
39	HPMPC (Cidofovir)	0	0	0	0	0	0	0	0	0	0	0	0	1	0
40	Hytrin	0	0	0	0	0	0	1	0	0	0	0	0	0	1

	A	N	O	P	Q	R	S	T	U	V	W	X	Y	Z	AA
1															
2		Carbonyl										Other N groups			
3		Ketone	Aldehyde	Enone	Ester	1o amide	2o amide	3o amide	Anhydride	Epoxide	Thioester	Oxime	Oxazolidinone	Urea	Guanidine
41	Imdur (Isosorbide mononitrate)	0	0	0	0	0	0	0	0	0	0	0	0	0	0
42	Iodixanol	0	0	0	0	0	4	2	0	0	0	0	0	0	0
43	Iohexol	0	0	0	0	0	2	1	0	0	0	0	0	0	0
44	Iopamidol	0	0	0	0	0	3	0	0	0	0	0	0	0	0
45	Isoflurane	0	0	0	0	0	0	0	0	0	0	0	0	0	0
46	Isradipine	0	0	0	2	0	0	0	0	0	0	0	0	0	0
47	Ivermectin	0	0	0	1	0	0	0	0	0	0	0	0	0	0
48	Ketoprofen	1	0	0	0	0	0	0	0	0	0	0	0	0	0
49	Klacid (Clarithromycin)	1	0	0	1	0	0	0	0	0	0	0	0	0	0
50	Levothyroxine	0	0	0	0	0	0	0	0	0	0	0	0	0	0
51	Lupron (Leuproreline)	0	0	0	0	0	10	0	0	0	0	0	0	0	1
52	Marcaine (Bupivacaine)	0	0	0	0	0	1	0	0	0	0	0	0	0	0
53	Meperidine	0	0	0	1	0	0	0	0	0	0	0	0	0	0
54	Meprobamate	0	0	0	0	0	0	0	0	0	0	0	0	0	0
55	Methohexital	0	0	0	0	0	0	0	0	0	0	0	0	2	
56	Metolazone	0	0	0	0	0	0	1	0	0	0	0	0	0	0
57	Mometasone furoate anhydrous	2	0	0	1	0	0	0	0	0	0	0	0	0	0
58	Mometasone furoate monohydrate	2	0	0	1	0	0	0	0	0	0	0	0	0	0
59	Nimbex (Cisatracurium besilate)	0	0	0	2	0	0	0	0	0	0	0	0	0	0
60	Nizatidine	0	0	0	0	0	0	0	0	0	0	0	0	0	0
61	Olanzapine	0	0	0	0	0	0	0	0	0	1	0	0	0	0
62	Oxis (Formoterol)	0	0	0	0	0	1	0	0	0	0	0	0	0	0
63	Paricalcitol (Zemplar)	0	0	0	0	0	0	0	0	0	0	0	0	0	0
64	Plendil (Felodipine)	0	0	0	2	0	0	0	0	0	0	0	0	0	0
65	Progesterone	1	0	1	0	0	0	0	0	0	0	0	0	0	0
66	Quinapril	0	0	0	1	0	1	1	0	0	0	0	0	0	0
67	Ranitidine	0	0	0	0	0	0	0	0	0	0	0	0	0	0
68	Roxithromycin	0	0	0	1	0	0	0	0	0	0	1	0	0	0
69	Salmeterol xinafoate	0	0	0	0	0	0	0	0	0	0	0	0	0	0
70	Sevelamer	0	0	0	0	0	0	0	0	0	0	0	0	0	0
71	Severane	0	0	0	0	0	0	0	0	0	0	0	0	0	0
72	Sumatriptan base	0	0	0	0	0	0	0	0	0	0	0	0	0	0
73	Tamsulosin	0	0	0	0	0	0	0	0	0	0	0	0	0	0
74	Venlafaxine	0	0	0	0	0	0	0	0	0	0	0	0	0	0
75	Warfarin	1	0	0	1	0	0	0	0	0	0	0	0	0	0

	A	AB	AC	AD	AE	AF	AG	AH	AI	AJ	AK	AL	AM	AN
1														
2		Other												
3		Ether	Sulfonamide	Sulfone	N-Oxide	Nitrile	Thiol	Thioether	Fluorine	Pyridine	Alkyl halide	Aryl halide	Alkene	Alkyl >5 C
4	Advicor (Niacin + Lovastatin)	1	0	0	0	0	0	0	0	0	0	0	0	1
5	Aluvia (Lopinavir + Ritonavir)	0	0	0	0	0	0	0	0	0	0	0	0	9
6	Androgel (Testosterone)	0	0	0	0	0	0	0	0	0	0	0	0	0
7	Atenolol	1	0	0	0	0	0	0	0	0	0	0	0	0
8	Bambec (Bambuterol)	0	0	0	0	0	0	0	0	0	0	0	0	0
9	Beclomethasone dipropionate	0	0	0	0	0	0	0	0	0	1	0		0
10	Beclomethasone dipropionate monohydrate	0	0	0	0	0	0	0	0	0	1	0	0	0
11	Betamethasone acetate	0	0	0	0	0	0	0	1	0	0	0	0	0
12	Betamethasone disodium phosphate	0	0	0	0	0	0	0	1	1	0	1	0	0
13	Blopress (candesartan cilexetil)	1	0	0	0	0	0	0	0	0	0	0	0	0
14	Brofen (Ibuprofen??)	0	0	0	0	0	0	0	0	0	0	0	0	0
15	Calcijex (Calcitriol)	0	0	0	0	0	0	0	0	0	0	0	3	2
16	Carisoprodol	0	0	0	0	0	0	0	0	0	0	0	0	0
17	Ciclesonide	2	0	0	0	0	0	0	0	0	0	0	0	0
18	ciclosporin	0	0	0	0	0	0	0	0	0	0	0	0	5
19	Citanest (Prilocaine)	0	0	0	0	0	0	0	0	0	0	0	0	0
20	Clarithromycin	6	0	0	0	0	0	0	0	0	0	0	0	0
21	Clobetasol propionate	0	0	0	0	0	0	0	1	0	1	0	0	0
22	Conholip	0	0	0	0	0	0	0	0	0	0	0	0	1
23	Cycloserine	0	0	0	0	0	0	0	0					
24	Deflox (Terezosin hydrochloride)	3	0	0	0	0	0	0	0	0	0	0	0	0
25	Dexamethosone dipropionate	0	0	0	0	0	0	0	1	0	0	0	0	0
26	Doxycycline hyclate	0	0	0	0	0	0	0	0	0	0	0	0	0
27	Doxycycline monohydrate	0	0	0	0	0	0	0	0	0	0	0	0	0
28	Epival (Sodium valproate)	0	0	0	0	0	0	0	0	0	0	0	0	0
29	Eprosartan (Teveten)	0	0	0	0	0	0	1	0	0	0	0	0	0
30	Fluticasone furaroate	1	0	0	0	0	0	0	3	0	0	0	0	0
31	Fluticasone propionate	0	0	0	0	0	0	0	3	0	0	0	0	0
32	Folic Acid	0	0	0	0	0	0	0	0	0	0	0	0	0
33	Furosemide	0	1	0	0	0	0	0	0	0	0	1	0	0
34	Gabapentin	0	0	0	0	0	0	0	0	0	0	0	0	0
35	Gadopentetate dimeglumine		0	0	0	0	0	0	0	0	0	0	0	0
36	Gadopentetate monomeglumine	0	0	0	0	0	0	0	0	0	0	0	0	0
37	gopten (Trandonapril)				0	0	0	0	0	0	0	0	0	0
38	Halobetasol	0	0	0	0	0	0	0	2	0	1	0	0	0
39	HPMPC (Cidofovir)	0	0	0	0	0	0	0	0	0	0	0	0	0
40	Hytrin	3	0	0	0	0	0	0	0	0	0	0	0	0

	A	AB	AC	AD	AE	AF	AG	AH	AI	AJ	AK	AL	AM	AN
1														
2		Other												
3		Ether	Sulfonamide	Sulfone	N-Oxide	Nitrile	Thiol	Thioether	Fluorine	Pyridine	Alkyl halide	Aryl halide	Alkene	Alkyl >5 C
41	Imdur (Isosorbide mononitrate)	2	0	0	1	0	0	0	0	0	0	0	0	0
42	Iodixanol	0	0	0	0	0	0	0	0	0	0	0	0	0
43	Iohexol	0	0	0	0	0	0	0	0	0	0	0	0	0
44	Iopamidol	0	0	0	0	0	0	0	0	0	0	0	0	0
45	Isoflurane	1	0	0	0	0	0	0	0	0	0	0	0	0
46	Isradipine	0	0	0	0	0	0	0	0	2	0	0	0	1
47	Ivermectin	9	0	0	0	0	0	0	0	0	0	0	0	0
48	Ketoprofen	0	0	0	0	0	0	0	0	0	0	0	0	0
49	Klacid (Clarithromycin)	6	0	0	0	0	0	0	0	0	0	0	0	0
50	Levothyroxine	1	0	0	0	0	0	0	0	0	0	4	0	0
51	Lupron (Leuproreline)	0	0	0	0	0	0	0	0	0	0	0	0	2
52	Marcaine (Bupivacaine)	0	0	0	0	0	0	0	0	0	0	0	0	0
53	Meperidine	0	0	0	0	0	0	0	0	0	0	0	0	0
54	Meprobamate	0	0	0	0	0	0	0	0	0	0	0	0	0
55	Methohexital	0	0	0	0	0	0	0	0	0	0	0	0	0
56	Metolazone	0	1	0	0		0	0	0	0	0	1	0	0
57	Mometasone furoate anhydrous	1	0	0	0	0	0	0	0	0	2	0	0	0
58	Mometasone furoate monohydrate	1	0	0	0	0	0	0	0	0	2	0	0	0
59	Nimbex (Cisatracurium besilate)	8	0	2	0	0	0	0	0	0	0	0	0	0
60	Nizatidine	0	0	0	0	0	0	1	0	0	0	0	0	0
61	Olanzapine	0	0	0	0	0	0	0	0	0	0	0	0	0
62	Oxis (Formoterol)	1	0	0	0	0	0	0	0	0	0	0	0	0
63	Paricalcitol (Zemplar)	0	0	0	0	0	0	0	0	0	0	0	3	0
64	Plendil (Felodipine)	0	0	0	0	0	0	0	0	0	0	2	2	0
65	Progesterone	0	0	0	0	0	0	0	0	0	0	0	0	0
66	Quinapril	1	0	0	0	0	0	0	0	0	0	0	0	0
67	Ranitidine	0	0	0	0	0	0	1	0	0	0	0	0	0
68	Roxithromycin	6	0	0	0	0	0	0	0	0	0	0	0	0
69	Salmeterol xinafoate	1	0	0	0	0	0	0	0	0	0	0	0	0
70	Sevelamer	0	0	0	0	0	0	0	0	0	0	0	0	0
71	Severane	1	0	0	0	0	0	0	0	0	0	0	0	0
72	Sumatriptan base	0	1	0	0	0	0	0	0	0	0	0	0	0
73	Tamsulosin	3	1	0	0	0	0	0	0	0	0	0	0	0
74	Venlafaxine	1	0	0	0	0	0	0	0	0	0	0	0	0
75	Warfarin	0	0	0	0	0	0	0	0	0	0	0	0	0

	A	AO	AP	AQ	AR	AS	AT	AU	AV	AW	AX	AY	AZ
1													
2								Organic framework					
3		Phosphonate	Hydrozone	Other	Phosphate	Carbamate	Nitro	Nitrate	Steroid	Hormone	O-heterocyclic	N-heterocyclic	S-heterocyclic
4	Advicor (Niacin + Lovastatin)	0	0	0	0	0	0	0	0	0	0	1	0
5	Aluvia (Lopinavir + Ritonavir)	0	0	0	0	0	0	0	0	0	0	0	0
6	Androgel (Testosterone)	0	0	0	0	0	0	0	1	0	0	0	0
7	Atenolol	0	0	0	0	0	0	0	0	0	0	0	0
8	Bambec (Bambuterol)	0	0	0	0	2	0	0	0	0	0	0	0
9	Beclomethasone dipropionate	0	0	0	0	0	0	0	1	0	0	0	0
10	Beclomethasone dipropionate monohydrate	0	0	0	0	0	0	0	1	0	0	0	0
11	Betamethasone acetate	0	0	0	0	0	0	0	1	0	0	0	0
12	Betamethasone disodium phosphate	1	1	1	1	0	0	0	1	0	0	0	0
13	Blopress (candesartan cilexetil)	0	0	0	0	0	0	0	0	0	0	2	0
14	Brofen (Ibuprofen??)	0	0	0	0	0	0	0	0	0	0	0	0
15	Calcijex (Calcitriol)	0	0	0	0	0	0	0	0	0	0	0	0
16	Carisoprodol	0	0	0	0	2	0	0	0	0	0	0	0
17	Ciclesonide	0	0	0	0	0	0	0	1	0	0	0	0
18	ciclosporin	0	0	0	0	0	0	0	0	0	0	0	0
19	Citanest (Prilocaine)	0	0	0	0	0	0	0	0	0	0	0	0
20	Clarithromycin	0	0	0	0	0	0	0	0	0	0	0	0
21	Clobetasol propionate	0	0	0	0	0	0	0	1	0	0	0	0
22	Conholip	0	0	0	0	0	0	0	0	0	0	0	0
23	Cycloserine												
24	Deflox (Terezosin hydrochloride)	0	0	0	0	0	0	0	0	0	0	1	0
25	Dexamethosone dipropionate	0	0	0	0	0	0	0	1	0	0	0	0
26	Doxycycline hyclate	0	0	0	0	0	0	0	0	0	0	0	0
27	Doxycycline monohydrate	0	0	0	0	0	0	0	0	0	0	0	0
28	Epival (Sodium valproate)	0	0	0	0	0	0	0	0	0	0	0	0
29	Eprosartan (Teveten)	0	1	0	0	0	0	0	0	0	0	1	0
30	Fluticasone furaroate	0	0	0	0	0	0	0	1	0	0	0	0
31	Fluticasone propionate	0	0	0	0	0	0	0	1	0	0	0	0
32	Folic Acid	0	0	0	0	0	0	0	0	0	0	2	0
33	Furosemide	0	0	0	0	0	0	0	0	0	1	0	0
34	Gabapentin	0	0	0	0	0	0	0	0	0	0	0	0
35	Gadopentetate dimeglumine	0	0	0	0	0	0	0	0	0	0	0	0
36	Gadopentetate monomeglumine	0	0	0	0	0	0	0	0	0	0	0	0
37	gopten (Trandonapril)	0	0	0	0	0	0	0	0	0	0	1	0
38	Halobetasol	0	0	0	0		0	0	1	0	0	0	0
39	HPMPC (Cidofovir)	1	0	0	0	0	0	0	0	0	0	1	0
40	Hytrin	0	0	0	0	0	0	0	0	0	0	1	0

	A	AO	AP	AQ	AR	AS	AT	AU	AV	AW	AX	AY	AZ
1													
2									Organic framework				
3		Phosphonate	Hydrozone	Other	Phosphate	Carbamate	Nitro	Nitrate	Steroid	Hormone	O-heterocyclic	N-heterocyclic	S-heterocyclic
41	Imdur (Isosorbide mononitrate)	0	0	0	0	0	0	1	0	0	1	0	0
42	Iodixanol	0	0	0	0	0	0	0	0	0	0	0	0
43	Iohexol	0	0	0	0	0	0	0	0	0	0	0	0
44	Iopamidol	0	0	0	0	0	0	0	0	0	0	0	0
45	Isoflurane	0	0	0	0	0	0	0	0	0	0	0	0
46	Isradipine	0	0	0	0	0	0	0	0	0	0	2	0
47	Ivermectin	0	0	0	0	0	0	0	0	0	0	0	0
48	Ketoprofen	0	0	0	0	0	0	0	0	0	0	0	0
49	Klacid (Clarithromycin)	0	0	0	0	0	0	0	0	0	0	0	0
50	Levothyroxine	0	0	0	0	0	0	0	0	1	0	0	0
51	Lupron (Leuproreline)	0	0	0	0	0	0	0	0	0	0	2	0
52	Marcaine (Bupivacaine)	0	0	0	0	0	0	0	0	0	1	0	0
53	Meperidine	0	0	0	0	0	0	0	0	0	0	1	0
54	Meprobamate	0	0	0	0	2	0	0	0	0	0	0	0
55	Methohexital	0	0	0	0	0	0	0	0	0	0	1	0
56	Metolazone	0	0	0	0	0	0	0	0	0	0	1	0
57	Mometasone furoate anhydrous	0	0	0	0	0	0	0	1	0	0	0	0
58	Mometasone furoate monohydrate	0	0	0	0	0	0	0	1	0	0	0	0
59	Nimbex (Cisatracurium besilate)	0	0	0	0	0	0	0	0	0	0	1	0
60	Nizatidine	0	0	0	0	0	1	0	0	0	0	1	1
61	Olanzapine	0	0	0	0	0	0	0	0	0	0	1	0
62	Oxis (Formoterol)	0	0	0	0	0	0	0	0	0	0	0	0
63	Paricalcitol (Zemplar)	0	0	0	0	0	0	0	0	0	0	0	0
64	Plendil (Felodipine)	0	0	0	0	0	0	0	0	0	0	1	0
65	Progesterone	0	0	0	0	0	0	0	1	1	0	0	0
66	Quinapril	0	0	0	0	0	0	0	0	0	0	2	0
67	Ranitidine	0	0	0	0	0	1	0	0	0	1	0	0
68	Roxithromycin	0	0	0	0	0	0	0	0	0	0	0	0
69	Salmeterol xinafoate	0	0	0	0	0	0	0	0	0	0	0	0
70	Sevelamer	0	0	0	0	0	0	0	0	0	0	0	0
71	Severane	0	0	0	0	0	0	0	0	0	0	0	0
72	Sumatriptan base	0	0	0	0	0	0	0	0	0	0	1	0
73	Tamsulosin	0	0	0	0	0	0	0	0	0	0	0	0
74	Venlafaxine	0	0	0	0	0	0	0	0	0	0	0	0
75	Warfarin	0	0	0	0	0	0	0	0	0	1	0	0

	A	BA	BB	BC	BD	BE	BF	BG	BH	BI	BJ	BK	BL	BM
1														
2														
3		Long alkyl	Phenyl ring	Erythromycin deriv	Tetracycline	Macrocyclic	Macrolide	Benzodiazepine	Barbiturate	Water	Ethanol	HCl	Na+	Gd3+
4	Advicor (Niacin + Lovastatin)	0	0	0	0	0	0	0	0	0	0	0	0	0
5	Aluvia (Lopinavir + Ritonavir)	0	0	0	0	0	0	0	0	0	0	0	0	0
6	Androgel (Testosterone)	0	0	0	0	0	0	0	0	0	0	0	0	0
7	Atenolol	0	1	0	0	0	0	0	0	0	0	0	0	0
8	Bambec (Bambuterol)	0	1	0	0	0	0	0	0	0	0	0	0	0
9	Beclomethasone dipropionate	0	0	0	0	0	0	0	0	0	0	0	0	0
10	Beclomethasone dipropionate monohydrate	0	0	0	0	0	0	0	0	1	0	0	0	0
11	Betamethasone acetate	0	0	0	0	0	0	0	0		0	0	0	0
12	Betamethasone disodium phosphate	0	0	0	0	0	0	0	0	0	0	0	2	0
13	Blopress (candesartan cilexetil)	0	3	0	0	0	0	0	0	0	0	0	0	0
14	Brofen (Ibuprofen??)	0	1	0	0	0	0	0	0	0	0	0	0	0
15	Calcijex (Calcitriol)	0	0	0	0	0	0	0	0	0	0	0	0	0
16	Carisoprodol	0	0	0	0	0	0	0	0	0	0	0	0	0
17	Ciclesonide	0	0	0	0	0	0	0	0	0	0	0	0	0
18	ciclosporin	0	0	0	0	1	0	0	0	0	0	0	0	0
19	Citanest (Prilocaine)	0	1	0	0	0	0	0	0	0	0	0	0	0
20	Clarithromycin	0	0	0	0	0	1	0	0	0	0	0	0	0
21	Clobetasol propionate	0	0	0	0	0	0	0	0	0	0	0	0	0
22	Conholip	0	0	0	0	0	0	0	0	0	0	0	0	0
23	Cycloserine	0	0	0	0	0	0	0	0	0	0	0	0	0
24	Deflox (Terezosin hydrochloride)	0	1	0	0	0	0	0	0	2	0	0	0	0
25	Dexamethosone dipropionate	0	0	0	0	0	0	0	0	0	0	0	0	0
26	Doxycycline hyclate	0	0	0	1	0	0	0	0	0.5	0.5	1	0	0
27	Doxycycline monohydrate	0		0	1	0	0	0	0	1	0	0	0	0
28	Epival (Sodium valproate)	0	0	0	0	0	0	0	0	0	0	0	1	0
29	Eprosartan (Teveten)	0	0	0	0	0	0	0	0	0	0	0	0	0
30	Fluticasone furaroate	0	0	0	0	0	0	0	0	0	0	0	0	0
31	Fluticasone propionate	0	0	0	0	0	0	0	0	0	0	0	0	0
32	Folic Acid	0	0	0	0	0	0	0	0	0	0	0	0	0
33	Furosemide	0	1	0	0	0	0	0	0	0	0	0	0	0
34	Gabapentin	0	0	0	0	0	0	0	0	0	0	0	0	0
35	Gadopentetate dimeglumine	0	0	0	0	0	0	0	0	0	0	0	0	1
36	Gadopentetate monomeglumine	0	0	0	0	0	0	0	0	0	0	0	0	1
37	gopten (Trandonapril)		1	0	0	0	0	0	0	0	0	0	0	0
38	Halobetasol	0	0	0	0	0	0	0	0	0	0	0	0	0
39	HPMPC (Cidofovir)	0	0	0	0	0	0	0	0	0	0	0	0	0
40	Hytrin	0	1	0	0	0	0	0	0	0	0	0	0	0

	A	BA	BB	BC	BD	BE	BF	BG	BH	BI	BJ	BK	BL	BM
1														
2														
3		Long alkyl	Phenyl ring	Erythromycin deriv	Tetracycline	Macrocyclic	Macrolide	Benzodiazepine	Barbiturate	Water	Ethanol	HCl	Na+	Gd3+
41	Imdur (Isosorbide mononitrate)	0	0	0	0	0	0	0	0	0	0	0	0	0
42	Iodixanol	0	0	0	0	0	0	0	0	0	0	0	0	0
43	Iohexol	0	0	0	0	0	0	0	0	0	0	0	0	0
44	Iopamidol	0	0	0	0	0	0	0	0	0	0	0	0	0
45	Isoflurane	0	0	0	0	0	0	0	0	0	0	0	0	0
46	Isradipine	0	0	0	0	0	0	0	0	0	0	0	0	0
47	Ivermectin	0	0	0	0	0	1	0	0	0	0	0	0	0
48	Ketoprofen	0	0	0	0	0	0	0	0	0	0	0	0	
49	Klacid (Clarithromycin)	0	0	0	0	0	1	0	0	0	0	0	0	0
50	Levothyroxine	0	2	0	0	0	0	0	0	0	0	0	0	0
51	Lupron (Leuproreline)	0	2	0	0	0	0	0	0	0	0	0	0	0
52	Marcaine (Bupivacaine)	0	1	0	0	0	0	0	0	0	0	0	0	0
53	Meperidine	0	1	0	0	0	0	0	0	0	0	0	0	0
54	Meprobamate	0	0	0	0	0	0	0	0	0	0	0	0	0
55	Methohexital	0	0	0	0	0	0	0	1					
56	Metolazone	0	2	0	0	0	0	0	0	0	0	0	0	0
57	Mometasone furoate anhydrous	0	0	0	0	0	0	0	0	0	0	0	0	0
58	Mometasone furoate monohydrate	0	0	0	0	0	0	0	0	1	0	0	0	0
59	Nimbex (Cisatracurium besilate)	0	0	0	0	0	0	0	0	0	0	0	0	0
60	Nizatidine	0	0	0	0	0	0	0	0	0	0	0	0	0
61	Olanzapine	0	0	0	0	0	0	1	0	0	0	0	0	0
62	Oxis (Formoterol)	0	2	0	0	0	0	0	0	0	0	0	0	0
63	Paricalcitol (Zemplar)	1	0	0	0	0	0	0	0	0	0	0	0	0
64	Plendil (Felodipine)	0	1	0	0	0	0	0	0	0	0	0	0	0
65	Progesterone	0	0	0	0	0	0	0	0	0	0	0	0	0
66	Quinapril	0	0	0	0	0	0	0	0	0	0	0	0	0
67	Ranitidine	0	0	0	0	0	0	0	0	0	0	0	0	0
68	Roxithromycin	0	0	1	0	0	1	0	0	0	0	0	0	0
69	Salmeterol xinafoate	1	0	0	0	0	0	0	0	0	0	0	0	0
70	Sevelamer	0	0	0	0	0	0	0	0	0	0	0	0	0
71	Severane	0	0	0	0	0	0	0	0	0	0	0	0	0
72	Sumatriptan base	0	0	0	0	0	0	0	0	0	0	0	0	0
73	Tamsulosin	0	0	0	0	0	0	0	0	0	0	0	0	0
74	Venlafaxine	0	0	0	0	0	0	0	0	0	1	0	0	0
75	Warfarin	0	2	0	0	0	0	0	0	0	0	0	0	0

Database 2: Physiochemical Properties

	A	B	C	D	E	F	G	H	I	J	K	L	M
1	Structure Name	Dermatological	Nasal and inhalation	Injectable	Antibiotic	API	Chemical Formula	Exact Mass	Molecular weight	C	O	F	H
2	Beclomethasone dipropionate monohydrate	1	1	0	0	0	C28H39CLO 8	538.23	539.06	63.39	23.74	0	7.29
3	Fluticasone propionate	1	1	0	0	0	C22H31F3O5 S	500.18	500.57	59.99	15.9	11.39	6.24
4	Mometasone furoate anhydrous	1	1	0	0	0				0	0	0	0
5	Mometasone furoate monohydrate	0	1	0	0	0	C27H30Cl2O 6	520.14	521.43	62.19	18.4	0	5.8
6	Sumatriptan base	0	1	0	0	0	C14H21N3O2 S	295.14	295.4	556.92	10.83	0	7.17
7	Ciclesonide	0	1	0	0	0	C33H46O7	554.32	554.71	71.45	20.19	0	5.59
8	Fluticasone furoate	0	1	0	0	0	C30H33F3O 7S	594.19	594.19	60.59	18.83	9.58	5.59
9	Salmeterol xinafoate	0	1	0	0	0	C36H45NO7	603.32	603.75	71.62	18.55	0	7.51
10	Beclomethasone dipropionate	0	0	0	0	0	C28H37F37 Cl07	521.042	0	0	0	0	0
11	Clobetasol propionate	0	0	0	0	0	C25H32ClF O5	466.19	466.97	64.3	17.13	4.07	6.91
12	Dexamethasone dipropionate	1	0	0	0	0	C28H37FO7	504.25	504.59	66.65	22.2	3.77	7.39
13	Halobetasol	1	0	0	0	0	C24H29ClF2 O5	470.93	470.93	61.21	16.99	8.07	6.21
14	Betamethasone acetate	0	0	1	0	0	C24H31FO6	434.21	434.5	66.34	22.09	4.37	7.19
15	Betamethasone disodium phosphate	0	0	1	0	0	C22H28FNa 3O8P+	539.12	539.39	48.99	23.73	3.52	5.23

	A	B	C	D	E	F	G	H	I	J	K	L	M
1	Structure Name	Dermatological	Nasal and inhalation	Injectable	Antibiotic	API	Chemical Formula	Exact Mass	Molecular weight	C	O	F	H
16	Doxcycline hyclate	0	0	0	1	0	C24H33ClN2O10	544.18	544.98	52.89	29.36	0	6.1
17	Doxcycline monohydrate	0	0	0	1	0	C22H26N2O9	462.16	462.45	57.14	31.14	0	5.67
18	Minocycline hydrochloride	0	0	0	1	0	C23H28ClN3O7	493.16	493.94	55.93	2.67	0	5.71
19	Roxithromycin	0	0	0	1	0	C41H76N2O15	836.52	837.05	58.83	28.67	0	9.15
20	Iodixanol	0	0	1	0	0	C35H44I6N6O15	1549.71	1550.18	27.12	15.48	0	2.86
21	Iopamidol	0	0	1	0	0	C17H22I3N3O8	776.85	777.09	26.28	16.47	0	2.85
22	Iohexol	0	0	1	0	0	C19H26I3N3O9	820.88	821.14	27.79	17.54	0	3.19
23	Gadopentetate dimeglumine	0	0	1	0	0	C28H54GdN5O20	937.99988		0	0	0	0
24	Gadopentetate monomeglumine	0	0	1	0	0	C14H20GdN3O10	547.57269		0	0	0	0
25	Ivermectin human grade	0	0	0	0	1	C48H74O14	874.51	875.09	65.88	25.6	0	8.52
26	Tamsulosin	0	0	0	0	1	C20H28N2O5S	408.17	408.17	58.8	19.58	0	6.91
27	Imdur					1	C6H9NO6	191.13879	191.138794			0	
28	Marcaine			1			C18H28N2O	288.22	288.43	74.96	5.55	0	9.78
29	Citanest			1			C13H21ClN2O	256.13	256.77	60.81	6.23	0	8.24

	A	B	C	D	E	F	G	H	I	J	K	L	M
1	Structure Name	Dermatological	Nasal and inhalation	Injectable	Antibiotic	API	Chemical Formula	Exact Mass	Molecular weight	C	O	F	H
30	Bambec		1				C18H30ClN3O5	403.19	403.9	53.53	19.81	0	7.49
31	Oxis		1				C23H28N2O8	344.17	344.4	66.26	18.58	0	7.02
32	Plendil					1	C18H19CL2NO4	383.25	384.25	56.26	16.66	0	4.98
33	Furosemide					1	C12H11ClN2O5S		330.74414	330.01	43.58	0	3.35
34	Atenolol					1	C14H22N2O3	266.16	266.34	63.13	18.02	0	8.33
35	Warfarin					1	C19H16O4	308.1	308.33	74.01	20.76	0	5.23
36	Meperidine					1	C15H21NO2	247.16	247.33	72.84	12.94	0	8.56
37	Metronazole				1		C6H16ClN3O3S	365.06	365.83	52.53	13.12	0	4.41
38	Ranitidine					1	C13H22N4O3S	314.14	314.4	49.66	15.27	0	7.05
39	Clarithromycin				1		C38H69NO13	747.48	747.95	61.02	27.81	0	9.3
40	Aluvia	1					C37H48N4O5	628.36	628.8	70.67	12.72	0	7.69

	A	B	C	D	E	F	G	H	I	J	K	L	M
1	Structure Name	Dermatological	Nasal and inhalation	Injectable	Antibiotic	API	Chemical Formula	Exact Mass	Molecular weight	C	O	F	H
41	Advicor					1	C30H41NO7		527.648987				
42	Deflox					1	C19H25N5O4	387.19	387.43	58.9	16.52	0	6.5
43	Levothyroxine					1	C15H11I4NO4	776.69	776.869995	23.19	8.24	0	1.43
44	Epival					1	C16H31NaO4		310.404694				
45	Gengaf					1	C26H111N11O12	1201.84	1202.61	61.92	15.96	0	9.3
46	Androgel	1					C19H28O2	288.21	288.42	79.12	11.09	0	9.78
47	Brofen					1	C13H18O2	206.13	206.28	75.69	51.51	0	8.8
48	Blopress					1	C24H20N6O3	440.16	440.45	65.45	10.9	0	4.58
49	Calcijex			1			C27H44O3	416.33	416.64	77.83	11.52	0	10.64
50	Venlafaxine					1	C17H28ClNO2	313.18	313.86	65.05	10.2	0	8.99
51	Ketoprofen					1	C16HO3	254.09	254.28	75.57	18.88	0	5.55
52	Quinapril					1	C25H30N2O5	438.22	438.52	68.47	18.24	0	6.9
53	Cycloserine					1	C3H6N2O2	102.04	102.09	35.29	31.34	0	5.92
54	Olanzapine			1			C17H20N4S	312.14	312.43	65.35	0	0	6.45

	A	B	C	D	E	F	G	H	I	J	K	L	M
1	Structure Name	Dermatological	Nasal and inhalation	Injectable	Antibiotic	API	Chemical Formula	Exact Mass	Molecular weight	C	O	F	H
55	HPMPC	1					C8H14N3O6P	279.06	279.19	34.42	34.38	0	5.05
56	Nizatidine					1	C12H21N5O2S2	331.11	331.46	43.48	9.65	0	6.39
57	Gabapentin					1	C9H17NO2	171.13	171.24	63.13	18.69	0	10.01
58	Meprobamate					1	C9H18N2O4	218.13	218.25	49.53	29.32	0	8.31
59	Folic Acid					1	C19H19N7O6	441.14	441.4	51.7	21.75	0	4.34
60	Methohexital			1			C14H18N2O3	286.13	286.3	58.73	16.76	0	6.69
61	Zemplar			1			C27H44O3	416.33	416.64	77.83	11.52	0	10.64
62	Teveten					1	C23H24NO4S	520.13	520.62	55.37	21.51	0	5.42
63	Sevorane		1				C4H3F7O	200.01	200.05	24.01	8	66.48	1.51
64	Prometrium					1	C21H30O2	314.22	314.46	80.21	10.18	0	9.62
65	Nimbex			1			C65H82N2O18S2	1242.5	1243.48	62.78	23.16	0	6.65

	A	B	C	D	E	F	G	H	I	J	K	L	M
1	Structure Name	Dermatological	Nasal and inhalation	Injectable	Antibiotic	API	Chemical Formula	Exact Mass	Molecular weight	C	O	F	H
66	Aluvia					1	C37H48N4O5	628.36	628.8	70.67	12.72	0	7.69
67	Conholip					1	C20H21ClO4		360.831299				
68	Isoflurane		1				C3H2ClF5O	183.97	184.49	19.22	8.67	51.49	1.09
69	Hytrin					1	C19H25N5O4	387.19	387.43	58.9	16.52	0	6.5
70	Gopten					1	C24H34N2O5	430.25	430.54	66.95	18.58	0	7.96
71	Klacid					1	C38H69NO13	747.48	747.95	61.02	27.81	0	9.3
72	Lupron					1	C61H88N16O14	1208.65	1209.4	58.59	15.88	0	7
73	Isradipine					1	C19H21N3O5		371.387085				
74	Selelamer					1	C6H13Cl2NO		186.079498				

	N	O	P	Q	R	S	T	U	V	W	X	Y	Z
1	S	N	P	Na	I	Cl	Elemental Analysis	Boiling Point [K]	Melting Point [K]	Critical Temp [K]	Critical Pressure [Bar]	Critical Volume [cm£/mol]	Gibbs Energy [kJ/mol]
2	0	0	0	0	0	6.58	C, 63.39, H, 7.29, Cl, 6.58, O, 23.74	678.2					
3	6.41	0	0	0	0	0	C,59.99,H,6.24,F,11.39,O,15.98,S,6.41	1139.5	771.08	976.44	12.54	1348.5	-918.11
4	0	0	0	0	0	0							
5	0	0	0	0	0	13.6	C,62.19,H,5.80,Cl,13.60,O,18.41	1248.4	877.28	1057.14	13.94	1386.5	-357.68
6	10.85	14.22	0	0	0	0	C,556.92,H,7.17,N,14.22,O,10.83,S,10.85	454.08	191.15				
7	0	0	0	0	0	0	C,71.45, H, 8.36,O, 20.19	1340.13	868.87	1098.37	10.01	1617.5	-375.72
8	5.39	0	0	0	0	0	C,60.59,H,5.59,F,9.58,O,18.83,S,5.39	1270.14	864.69	1063.2	10.21	1570.5	-1041.13
9	0	2.32	0	0	0	0	C71.62,H,7.51,N,2.32,O,18.55						
10	0	0	0	0	0	0							
11	0	0	0	0	0	7.59	C,64.30, H, 6.91, Cl, 7.59,F, 4.07, O, 17.13	1114.28	769.66	976.38	13.18	1308.5	-565.83
12	0	0	0	0	0	0	C,66.65, H, 7.39, F, 3.77, O, 22.20	1203.71	815.88	1013.35	11.1	1453.5	-839.01
13	0	0	0	0	0	7.53	C,61.21,H,6.21,Cl,7.53,F,8.07,O,16.99	1081.73	758.26	957.88	13.11	1277.5	-764.67
14	0	0	0	0	0	0	C,66.34, H, 7.19, F, 4.37,O, 22.09	1146	789.29	964.9	15.62	1222.5	-699.14
15	0	0	5.74	12.79	0	0	C,48.99,H, 5.23, F, 3.52, Na, 12.79, O, 23.73, P,5.74						

	N	O	P	Q	R	S	T	U	V	W	X	Y	Z
1	S	N	P	Na	I	Cl	Elemental Analysis	Boiling Point [K]	Melting Point [K]	Critical Temp [K]	Critical Pressure [Bar]	Critical Volume [cm£/mol]	Gibbs Energy [kJ/mol]
16	0	5.14	0	0	0	6.51	C,52.89,H,6.10,Cl,6.51,N,5.14,O,29.36						
17	0	6.06	0	0	0	0	C57.14,H5.67,N6.06,O,31.14						
18	0	8.51	0	0	0	7.18	C,55.93,H,5.71,Cl,7.18,N,8.51,O,2.67						
19	0	3.35	0	0	0	0	C,58.83,H,9.15,N,3.35,O,28.67						
20	0	5.42	0	0	49.12	0	C,27.12,H,2.86,I,49.12,N,5.42,O,15.48						
21	0	5.41	0	0	48.99	0	C,26.28,H2.85,I,48.99,N,5.41,O,16.47	1753.87	1301.78	1659.49	29.35	1304.5	-537.78
22	0	5.12	0	0	46.36	0	C,27.79,H,3.19,I,46.36,N,5.12,O,17.54	1833.21	1301.33	1870.95	25.13	1431.5	-616.82
23	0	0	0	0	0	0							
24	0	0	0	0	0	0							
25	0	0	0	0	0	0	C,65.88,H,8.52,O,25.60						
26	7.85	6.86	0	0	0	0	C,58.80,H,6.91,N,6.86,O,19.58,S,7.85	1087.71					
27	0		0	0	0	0		364.488	71.5	0	0	0	0
28	0	9.71	0	0	0	0	C,74.96, H, 9.78, N, 9.71, O, 5.5	800.71	553.16	934.64	18	922.5	282.08
29	0	10.91	0	0	0	13.81	C,60.81, H, 8.24, CL, 13.81, N, 10.91, O, 6.23	361.6					

	N	O	P	Q	R	S	T	U	V	W	X	Y	Z
1	S	N	P	Na	I	Cl	Elemental Analysis	Boiling Point [K]	Melting Point [K]	Critical Temp [K]	Critical Pressure [Bar]	Critical Volume [cm£/mol]	Gibbs Energy [kJ/mol]
30	0	10.4	0	0	0	8.78	C,53.53, H, 7.49, Cl, 8.78, N, 10.40, O, 19.81	500.865	154.12				
31	0	8.13	0	0	0	0	C, 66.26, H, 7.02, N, 8.13, O, 18.58	1040.98	693.36	945.38	23.34	1011.5	-7.4
32	0	3.65	0	0	0	18.45	C, 56.26, O, 16.66, H, 4.98, N, 3.65, Cl, 18.45	925.72	652.09	907.19	16.43	1027.5	-317.21
33	9.69	8.47	0	0	0	10.72	C, 43.58, H, 3.35, Cl, 10.72, N, 8.47, O, 24.19, S, 9.69	1044.31	211.61				
34	0	10.52	0	0	0	0	C, 63.13, H, 8.33, N, 10.52, O, 18.02	841.87	524.88	887.27	24.46	806.5	-50
35	0	0	0	0	0	0	C, 74.01, H, 5.23, O, 20.76	912.75	558.33	958.8	23.05	880.5	-163
36	0	5.66	0	0	0	0	C, 72.84, H, 8.56, N, 5.66, O, 12.94	655.66	394.33	809.83	22.7	750.5	19.3
37	8.76	11.49	0	0	0	9.69	C, 52.53, H, 4.41, Cl, 9.69, N, 11.49, O, 13.12, S, 8.76	1057.16	237.64				
38	10.2	17.82	0	0	0	0	C, 49.66, H, 7.05, N, 17.82, O, 15.27, S, 10.2	437.13					
39	0	1.87	0	0	0	0	C, 61.02, H, 9.3, N, 1.87, O, 27.81	805.478					
40	0	8.91	0	0	0	0	C, 70.67, H, 7.69, N, 8.91, O, 112.72	1626.27	1108.26	1535.32	9.56	1863.5	346.16

	N	O	P	Q	R	S	T	U	V	W	X	Y	Z
1	S	N	P	Na	I	Cl	Elemental Analysis	Boiling Point [K]	Melting Point [K]	Critical Temp [K]	Critical Pressure [Bar]	Critical Volume [cm£/mol]	Gibbs Energy [kJ/mol]
41													
42	0	18.08	0	0	0	0	C, 58.90, H, 6.50, N, 18.08, O, 16.52	1075.94	854.39	1034.08	23.05	1030.5	399.62
43	0	1.8	0	0	65.34	0	C, 23.19, H, 1.43, I, 65.34, N, 1.80, O, 8.24	1314.26	968.8	1144.49	28.51	1092.5	-54.95
44								220					
45	0	12.81	0	0	0	0	C, 61.92, H, 9.30, N, 12.81, O, 15.96	1293.782					
46	0	0	0	0	0	0	C, 79.12, H, 9.78, O, 11.09	837.91	539.19	850.34	20.2	893.5	26.12
47	0	0	0	0	0	0	C, 75.69, H, 8.80, O, 15.51	673.33	405.31	789.46	23.91	667.5	-187.43
48	0	19.08	0	0	0	0	C, 65.45, H, 4.58, N, 19.08, O, 10.90	755.37	331.89				
49	0	0	0	0	0	0	C, 77.83, H, 10.64, O, 11.52	1139.41	645.95	966.71	11.8	1356.5	-0.86
50	0	4.46	0	0	0	11.3	C, 65.05, H, 8.99, CL, 11.30, N, 4.46, O, 10.20		118.8				
51	0	0	0	0	0	0	C, 75.57, H, 5.55, O, 18.88	822.96	530.47	879.8	26.38	739.5	-176.24
52	0	6.39	0	0	0	0	C, 68.47, H, 6.90, N, 6.39, O, 18.24	1156.01	762.3	1049.32	14.17	1273.5	-152.35
53	0	27.44	0	0	0	0	C, 35.29, H, 5.92, N, 27.44, O, 31.34	478.5	353.43	785.05	63.8	251.5	-24.07
54	10.26	17.93	0	0	0	0	C, 65.35, H, 6.45, N, 17.93, S, 10.26		188.52				

	N	O	P	Q	R	S	T	U	V	W	X	Y	Z
1	S	N	P	Na	I	Cl	Elemental Analysis	Boiling Point [K]	Melting Point [K]	Critical Temp [K]	Critical Pressure [Bar]	Critical Volume [cm£/mol]	Gibbs Energy [kJ/mol]
55	0	15.05	11.09	0	0	0	C, 34.42, H, 5.05, N, 15.05, O, 34.38, P, 11.09		90.27				
56	19.35	21.13	0	0	0	0	C, 43.48, H, 6.39, N, 21.13, O, 9.65, S, 19.35						
57	0	8.18	0	0	0	0	C, 63.13,H, 10.01, N, 8.18, O, 18.69	643.35	465.83	784.09	35.18	523.5	-233.6
58	0	12.84	0	0	0	0	C, 49.53, H, 8.31, N, 12.84, O, 29.32	663.79	444.29	755.19	26.6	638.5	-460.1
59	0	22.21	0	0	0	0	C, 51.70, H, 4.34, N, 22.21, O, 21.75	1569.82	1432.7	1297.98	36.25	1151.5	-28.24
60	0	9.78	0	8.03	0	0	C, 58.73, H, 6.69, N, 9.78, Na, 8.03, O, 16.76	545.36					
61	0	0	0	0	0	0	C, 77.83, H, 10.64, O, 11.52	1143.97	612.19	963.36	12.14	1346.5	23.84
62	12.32	5.38	0	0	0	0	C, 55.37, H, 5.42, N, 5.38, O, 21.51, S, 12.32	660.623	285.23				
63	0	0	0	0	0	0	C, 24.01, H, 1.51, F, 66.48, O, 8.00	301.53	150.54	383.12	28.05	375.5	-1482.63
64	0	0	0	0	0	0	C, 80.21, H, 9.62, O, 10.18	845.36	550.84	868.2	16.71	992.5	50.86
65	5.16	2.25	0	0	0	0	C, 62.78, H, 6.65, N, 2.25, O, 23.16, S, 5.16						

	N	O	P	Q	R	S	T	U	V	W	X	Y	Z
1	S	N	P	Na	I	Cl	Elemental Analysis	Boiling Point [K]	Melting Point [K]	Critical Temp [K]	Critical Pressure [Bar]	Critical Volume [cm³/mol]	Gibbs Energy [kJ/mol]
66	0	8.91	0	0	0	0	C, 70.67, H, 7.69, N, 8.91, O, 12.72	1626.27	1108.26	1535.32	9.56	1863.5	346.16
67								469.774	157.69				
68	0	0	0	0	0	19.22	C, 19.53, H, 1.09, Cl, 19.22, F, 51.49, O, 8.67	320.33	150.59	443.8	32.65	337.5	-1118.64
69	0	18.08	0	0	0	0	C, 58.90, H, 6.50, N, 18.08, O, 16.52	1075.94	854.39	1034.08	23.05	1030.5	399.62
70	0	6.51	0	0	0	0	C, 66.95, H, 7.96, N, 6.51, O, 18.58	1112.08	718.75	1018.37	13.42	1265.5	-234.71
71	0	1.87	0	0	0	0	C, 61.02, H, 9.3, N, 1.87, O, 27.81						
72	0	18.53	0	0	0	0	C, 58.59, H, 7, N, 18.53, O, 15.88	1720.5					
73								501.946					
74								116.1					

	AA	AB	AC	AD	AE	AF	AG	AH	AI	AJ	AK	AL	AM	AN
1	Log P	MR [cm3/mol]	Henry's Law	Heat of Form	tPSA	CLogP	CMR	ACD/Log P	ACD/Log D (pH5.5)	ACD/BCF (Ph5.5)	ACD/KOC (pH5.5)	H bond acceptors	Freely rotating bonds	Index of refraction
2	4.43							3.209	3.21	161.66	1325.93	8	9	
3	3.12	122.92	4.34	-1501.83	80.67	3.0326	12.5188	3.73	3.73	402.74	258.47	5	7	1.556
4								4.267	4.27	1030.54	4992.88	6	6	
5	3.21	133.63	14.45	-954.88	89.9	2.37052	13.3576	2.675	2.68	63.48	679.14	4	4	1.6
6			4.47 E-04		61.44	0.742999	8.1988	0.436	-2.59	1	1	5	5	1.61
7	3.97	152.9	16.55	-1257.55	99.13	5.87195	15.2391	6.13	6.13	26845.77	51496.1	7	7	1.576
8	3.86	144.92	4.31	-1717.17	95.97	3.61452	14.6547	4.203	4.2	921.07	4607.25	6	8	1.579
9								3.074	0.05	1	1.05	5	19	
10								4.073	4.07	733.01	3912.45	7	9	1.564
11	2.63	121.17	12.87	-11146.88	80.67	3.15848	12.0842	3.142	3.14	143.87	1219.77	5	6	1.56
12	2.49	132.14	14.92	-1510.34	106.97	2.26712	13.173	3.666	3.67	360.11	2352.34	7	9	1.55
13	2.08	116.59	12.7	-1336.53	80.67	1.9538	11.6359	2.947	2.95	102.28	955.48	5	6	1.551
14	0.95	113.3	13.98					2.654	2.65	61.25	661.98	6	6	1.571
15												8	6	

	AA	AB	AC	AD	AE	AF	AG	AH	AI	AJ	AK	AL	AM	AN
1	Log P	MR [cm3/mol]	Henry's Law	Heat of Form	tPSA	CLogP	CMR	ACD/Log P	ACD/Log D (pH5.5)	ACD/BCF (Ph5.5)	ACD/KOC (pH5.5)	H bond acceptors	Freely rotating bonds	Index of refraction
16														
17														
18														
19														
20								-4.069	-4.07	1	1	21	31	1.752
21	0.98	135.48	30.16	-1098.04	188.45	-2.23633	13.6883							
22	0.37	146.01	31.13	-1224.39	199.89	-2.44413	14.769							
23								0.046	-5.72	1	1	13	16	
24								0.046	-5.72	1	1	13	16	
25														
26	2.69	112.23	2.26		99.88	2.1672	10.9794	2.139	-0.73	1	1	7	11	1.553
27	-0.177	38.447	5.69E-14	0	0	0	0	-0.177	-0.18	1	19.08	7	3	1.543
28	3.86	89.94	9.43	-185.46	32.34	3.6912	8.8499	3.312	1.27	1.77	13.82	3	5	1.547
29								1.738	-0.33	1	1.78	3	5	0

	AA	AB	AC	AD	AE	AF	AG	AH	AI	AJ	AK	AL	AM	AN
1	Log P	MR [cm3/mol]	Henry's Law	Heat of Form	tPSA	CLogP	CMR	ACD/Log P	ACD/Log D (pH5.5)	ACD/BCF (Ph5.5)	ACD/KOC (pH5.5)	H bond acceptors	Freely rotating bonds	Index of refraction
30			4.14E-17					1.043	-2.01	1	1	8	9	1.533
31	1.67	96.96	20.43	-436.33	90.82	1.2604	9.6788	1.566	-1.4	1	1	6	10	
32	2.24	98.5	1.35E-11	-705.49	64.63	5.2968	9.9071	4.761	4.76	2440.56	9250.27	5	6	1.55
33	0.74	76.49	9.66		118.72	1.90019	7.6314	2.304	-0.1	1	1.67	7	5	1.658
34	0.22	74.62	16.25	-427.56	84.58	-0.108601	7.4783	0.335	-2.75	1	1	5	9	1.54
35	2.97	85.93	11.03	-427.13	63.6	2.9013	8.7182	3.129	2.09	12.83	109.47	4	5	1.635
36	2.64	71.89	6.4	-290.43	29.54	2.227	7.2429	2.185	-0.08	1	1.98	3	4	1.52
37	2.64	95.02	11.3		92.5	2.05579	9.3833	3.163	3.16	149.15	1251.65	6	2	1.629
38			7.14		88.34	0.670201	8.6512	-0.068	-2.73	1	1	7	10	1.558
39								2.805	0.35	1	2.82	14	12	1.526
40	4.56	182.98	30.06	-554.07	120	6.0946	18.2339	5.417	5.42	7698.33	21047.25	9	16	1.577

	AA	AB	AC	AD	AE	AF	AG	AH	AI	AJ	AK	AL	AM	AN
1	Log P	MR [cm3/mol]	Henry's Law	Heat of Form	tPSA	CLogP	CMR	ACD/Log P	ACD/Log D (pH5.5)	ACD/BCF (Ph5.5)	ACD/KOC (pH5.5)	H bond acceptors	Freely rotating bonds	Index of refraction
41								4.066	4.07	724.49	3879.84	5	8	
42	1.13	107.91	22.02	-180.9	101.98	2.18152	10.3025	-0.963	-1.95	1	1	9	4	
43	7.36	122.92	17.79	-261.03	92.78	3.50768	12.6805	4.719	2.21	6.98	27.03	5	7	1.795
44								2.719	1.95	11.78	123.33	2	5	
45								3.351	3.35	207.1	1582.17	23	16	1.468
46	3.31	84.29	6.84	-428.91	37.3	-0.11016	8.5194	3.179	3.18	153.38	1276.96	2	1	1.56
47	3.75	61.2	5.21	-447.42	37.3	3.679	6.124	3.502	2.38	20.4	144.58	2	4	1.519
48	4.72	128.17	7.42		111.24	5.43126	12.1219	4.652	1.41	1.16	4.64	9	7	1.719
49	4.49	126.14	4.9	-679.03	60.69	4.475	12.8549	5.632	5.63	11219.63	27578.06	3	9	1.547
50			2.04E-11					2.475	-0.51	1	1	3	6	1.544
51	3.31	72.7	9.06	-380.11	54.37	2.761	7.2795	2.911	1.62	4.93	46.9	3	4	1.592
52	3.17	121.81	19.06	-694.63	95.94	1.9111	12.2025	4.788	2.6	16.7	62.44	7	10	1.578
53	-1.7	23.07	4.13	-189.77	64.35	-1.192	2.3176	-2.988	-3.54	1	1	4	1	1.475
54			14.56		30.87	3.009	9.1981	3.076	0.42	1	2.5	4	1	1.709

	AA	AB	AC	AD	AE	AF	AG	AH	AI	AJ	AK	AL	AM	AN
1	Log P	MR [cm3/mol]	Henry's Law	Heat of Form	tPSA	CLogP	CMR	ACD/Log P	ACD/Log D (pH5.5)	ACD/BCF (Ph5.5)	ACD/KOC (pH5.5)	H bond acceptors	Freely rotating bonds	Index of refraction
55	998.56	59.82	2.43		145.68	-2.3908	6.1065	-1.466	-4.68	1	1	9	7	1.656
56			10.29		91.47	-0.1618	9.0353	1.179	-0.49	1	2.26	7	10	1.592
57	0.88	45.22	8.13	-476.01	63.32	-0.66	4.7317	1.083	-1.47	1	1	3	4	1.489
58	1.06	53.73	11.32	-804.82	104.64	0.915	5.4666	0.7	0.7	2	57.25	6	8	1.479
59	0.57		38.39	-582.69	211.42	-0.725222	10.9625	-0.99	-3.8	1	1	13	9	1.763
60			1.52E-13					2.013	2.01	19.85	295.15	5	4	1.505
61	4.52	127.99	4.81	-651.33	60.69	5.688	12.7029	5.899	5.9	17930.17	38574.72	3	8	1.609
62			5.55E-17					0.959	-0.36	1	1	6	10	1.628
63	2.24	23.78	-0.88	-1653.66	9.23	1.451	2.2942	2.498	2.5	46.59	544.23	1	2	1.266
64	3.78	92.44	5.58E+00	-430.54	34.14	0.485839	9.3296	3.827	3.83	476.94	2876.39	2	1	1.542
65								1.038	1.04	3.62	87.42	14	26	

	AA	AB	AC	AD	AE	AF	AG	AH	AI	AJ	AK	AL	AM	AN
1	Log P	MR [cm3/mol]	Henry's Law	Heat of Form	tPSA	CLogP	CMR	ACD/Log P	ACD/Log D (pH5.5)	ACD/BCF (Ph5.5)	ACD/KOC (pH5.5)	H bond acceptors	Freely rotating bonds	Index of refraction
66	4.56	182.98	30.06	-554.07	120	6.0946	18.2339	5.417	5.42	7698.33	21047.25	9	16	1.577
67			4.46E-09					5.801	5.8	15087.66	34091.39	4	7	1.547
68	2.47	23.96	1.26E-02	-1253.07	9.23	1.764	2.2908	2.118	2.12	23.96	338.11	1	2	1.301
69	1.13	107.91	2.20E+01	-180.9	101.98	2.18152	10.3025	0.797	-0.25	1	5.77	9	4	1.636
70	2.9	117.2	1.76E+01	-858.91	95.94	1.05352	11.8329	4.9	2.64	17	60.14	7	10	1.549
71								2.805	0.35	1	2.82	14	12	1.526
72								0.629				30	34	
73								3.728	3.73	400.94	2539.49	8	6	1.566
74								0.45	0.45	1.29	41.84	1	1	

AO	AP	AQ	AR	AS	AT	AU	AV	AW	AX	AY	AZ	BA	BB
Molar Volume cm	Surface Tension dyne/cm	Flash point °c	Boiling Point °c	ACD/ Log D (pH7.4)	ACD/BCF (pH7.4)	ACD/KOC (pH7.4)	H bond donors	Polar surface area A	Molar Refractivity cm	Polarizability 10^{-24} cm	Density g/cm³	Enthalpy of vapourisation kJ/mol	Vapour pressure mmHg@25°c
		364	678.2	3.21	161.66	1325.92	3	106.97				113.89	2.61E-21
377.027	48.062997	297.491	568.289	3.73	402.73	2548.45	1	105.97	121.148	48.027 10^{-24}	1.328	97.97	
		350.2	655.5	4.27	1030.53	4992.87	1	82.81					4.47E-18
316.548	55.51599	308.544	586.566	2.68	63.48	679.12	2	74.6	108.251	42.914	1.35	100.559	0
237.622	52.723	254.811	497.718	-1.45	1	1	2	73.58	82.359	32.65	1.243	76.575	0
436.998	51.861	209.975	664.979	6.13	26845.77	51496.1	1	99.13	144.517	57.291	1.237	111.927	0
401.799	52.534	330.909	623.546	4.2	921.06	4607.21	1	119.11	133.714	53.009	1.375	97.082	0
		318.5	603	1.15	1.53	13.42	4	40.16				94.33	2.16E-15
410.299	51.581001	335.142	630.546	4.07	733.01	3912.44	1	106.97	133.382	52.877	1.27	106.882	
364.135	48.914	297.905	568.973	3.14	143.87	1219.77	1	80.67	117.751	46.68	1.282	98.067	0
403.95	49.18999	318.585	603.169	3.67	360.11	2352.33	1	106.97	128.667	51.008	1.249	102.93	0
369.639	47.438999	298.944	570.691	2.95	102.28	955.48	1	80.67	117.848	46.718	1.312	98.309	0
334.034	52.93299	304.206	579.392	2.65	61.25	661.96	2	100.9	109.821	43.537	1.301	99.54	0
							4	156.83					

	AO	AP	AQ	AR	AS	AT	AU	AV	AW	AX	AY	AZ	BA	BB
1	Molar Volume cm	Surface Tension dyne/cm	Flash point °c	Boiling Point °c	ACD/ Log D (pH7.4)	ACD/BCF (pH7.4)	ACD/KOC (pH7.4)	H bond donors	Polar surface area A	Molar Refractivity cm	Polarizability 10^{-24} cm	Density g/cm^3	Enthalpy of vapourisation kJ/mol	Vapour pressure mmHg@25°c
16														
17														
18														
19														
20	675.385	89.597	710.3	1250.869	-4.07	1	1	13	339.09	275.645	109.274	2.295	195.443	0
21														
22														
23			389.9	721.1	-5.95	1	1	5	174.22				114.7	9.53E-23
24			389.9	721.1	-5.95	1	1	5	196.22				114.7	9.53E-23
25														
26	342.817	45.282	313.942	595.492	0.77	1.05	14.72	3	108.26	109.784	43.522	1.192	88.739	0
27	122.014	60.254002	174.236	364.488	-0.18	1	19.08	1	93.74	38.447	15.241	1.567	70.678	0
28	279.243	41.578999	209.878	423.422	2.92	77.82	606.39	1	32.34	88.62	35.132	1.033	67.775	0
29	0		134.3	361.6	1.32	4.7	80.05	2	23.55				60.74	2.05E-05

	AO	AP	AQ	AR	AS	AT	AU	AV	AW	AX	AY	AZ	BA	BB
1	Molar Volume cm	Surface Tension dyne/cm	Flash point °c	Boiling Point °c	ACD/ Log D (pH7.4)	ACD/BCF (pH7.4)	ACD/KOC (pH7.4)	H bond donors	Polar surface area A	Molar Refractivity cm	Polarizability 10^{-24} cm	Density g/cm³	Enthalpy of vapourisation kJ/mol	Vapour pressure mmHg@25°c
30	318.174	43.14199	256.714	500.865	-1.05	1	1	2	91.34	98.822	39.176	1.155	81.039	0
31			318.6	603.2	-0.07	1	3.93	4	51.24				94.35	2.12E-15
32	300.844	42.194	238.964	471.516	4.76	2444.58	9265.53	1	64.63	95.782	37.971	1.277	73.428	0
33	205.814	75.271004	305.854	582.118	-0.82	1	1	4	131.01	75.767	30.036	1.607	91.548	3.06E-11
34	236.659	45.019001	261.059	508.049	-1.76	1	1	4	84.58	74.257	29.438	1.125	81.95	7.69E-10
35	235.758	58.658001	188.828	515.155	0.33	1	1.89	1	63.6	84.447	33.477	1.308	82.854	1.16E-07
36	234.243	38.337002	111.636	328.866	1.62	7.25	98.88	0	29.54	71.266	28.252	104.85	57.13	8.43E-07
37	261.206	53.708	324.892	613.598	3.16	148.79	1248.59	3	36.796	92.819	36.796	1.401	91.064	3.55E-12
38	265.446	45.015999	218.169	437.13	-1.07	1	2.17	2	111.56	85.647	33.953	1.184	69.37	1.68E+02
39	631.905	48.707	440.937	805.478	2.06	14.31	143.48	4	182.91	194.005	76.91	1.184	133.363	
40	540.456	49.473	512.707	924.15	5.42	7698.34	21047.28	4	120	179.178	71.032	1.163	140.807	

	AO	AP	AQ	AR	AS	AT	AU	AV	AW	AX	AY	AZ	BA	BB
1	Molar Volume cm	Surface Tension dyne/cm	Flash point 'c	Boiling Point 'c	ACD/ Log D (pH7.4)	ACD/BCF (pH7.4)	ACD/KOC (pH7.4)	H bond donors	Polar surface area A	Molar Refractivity cm	Polarizability 10^{-24} cm	Density g/cm^3	Enthalpy of vapourisation kJ/mol	Vapour pressure mmHg@25'c
41			185.3	559.2	4.07	724.49	3879.83	1	61.83				96.69	7.81E-15
42			355.7	664.5	-101	1	6.42	2	80.26				97.72	1.59E-17
43	294..745	79.616997	302.339	576.306	1.68	2.07	8.01	4	92.78	125.446	49.731	2.636	90.781	
44			116.6	220	0.16	1	2	1	26.3				50.28	0.0435
45	1183.625	31.621	736.253	1293.782	3.35	207.1	1582.17	5	278.8	328.831	130.359	1.016	218.52	
46	256.96	44.490002	184.655	432.925	3.18	153.38	1276.96	1	37.3	83.113	32.949	1.122	79.521	1.71E-08
47	200.339	38.677999	216.702	319.643	0.58	1	2.3	1	37.3	60.776	24.093	1.03	59.252	0.000186
48	310.519	59.644001	754.756	755.37	0.54	1	1	2	118.81	122.545	48.581	1.418	115.396	1.79E-18
49	391.894	44.083	238.428	565.009	5.63	11219.63	27578.06	3	60.69	124.354	49.298	1.063	97.508	1.19E-12
50	261.697	41.147999	194.246	397.575	0.76	1	10.12	1	32.7	82.634	32.759	1.06	68.342	
51	212.246	49.770999	228.793	431.316	-0.16	1	1	1	54.37	71.795	28.462	1.198	72.408	
52	360.125	52.294998	35.149	661.974	1.24	1	2.68	2	95.94	119.511	47.378	1.218	102.322	324.1
53	79.861	40.884983		281.12	-3	1	1	3	64.35	22.472	8.909	1.278		
54	235.993	52.015999	241.697	476.035	2.35	23.98	210.04	1	59.11	92.156	36.534	1.324	73.967	188.52

	AO	AP	AQ	AR	AS	AT	AU	AV	AW	AX	AY	AZ	BA	BB
1	Molar Volume cm	Surface Tension dyne/cm	Flash point °c	Boiling Point °c	ACD/ Log D (pH7.4)	ACD/BCF (pH7.4)	ACD/KOC (pH7.4)	H bond donors	Polar surface area A	Molar Refractivity cm	Polarizability 10^{-24} cm	Density g/cm^3	Enthalpy of vapourisation kJ/mol	Vapour pressure mmHg@25°c
55	1.58	90.557999	322.419	609.508	-5.41	1	1	5	155.49	58.293	23.109	1.76	103.841	90.27
56	265.354	51.514	242.984	478.162	0.98	2.92	65.63	2	139.55	89.823	35.608	1.249	74.221	
57	161.825	47.09	143.967	314.438	-1.42	1	1	3	63.32	46.696	18.512	1.058	61.095	2.94E-10
58	191.464	43.902	229.739	434.212	0.7	2	57.25	4	104.64	54.331	21.538	1.14	69.029	46.59
59	261.361	80.670009		846.45	-5.72	1	1	7	208.99	107.815	42.741	1.689		6.17E-20
60	235.572	39.894001		545.36	1.88	14.52	215.84	1	66.48	69.82	27.679	1.113		1.08E-11
61	371.436	54.659	238.344	564.843	5.9	17930.17	38574.72	3	60.69	128.646	50.999	1.122	97.485	8.61E-14
62	335.336	51.091999	353.332	660.623	-0.63	1	1	2	120.66	118.944	47.153	1.266	102.136	3.35E-15
63	139.532	13.027	-11.446	49.472	2.5	46.59	544.23	0	9.23	23.362	9.261	1.434	28.084	3.11E+02
64	288.952	41.171001	166.683	447.151	3.83	476.94	2876.39	0	34.14	90.955	36.057	1.088	70.544	2.69E-06
65					1.04	3.62	87.42		126.44					

	AO	AP	AQ	AR	AS	AT	AU	AV	AW	AX	AY	AZ	BA	BB
1	Molar Volume cm	Surface Tension dyne/cm	Flash point °c	Boiling Point °c	ACD/ Log D (pH7.4)	ACD/BCF (pH7.4)	ACD/KOC (pH7.4)	H bond donors	Polar surface area A	Molar Refractivity cm	Polarizability 10^{-24} cm	Density g/cm^3	Enthalpy of vapourisation kJ/mol	Vapour pressure mmHg@25°c
66	540.456	49.473	512.707	924.15	5.42	7698.34	21047.28	4	120	179.178	71.032	1.163		
67	306.449	40.981998	165.356	469.774	5.8	15087.66	34091.39	0	52.6	97.116	38.5	1.177	73.22	
68	123.843	15.828	10.643	48.49	2.12	23.96	338.11	0	9.23	23.244	9.215	1.49	28.001	322.98801
69	290.672	64.138	355.663	664.477	0.74	2.1	57.08	2	103.04	104.264	41.334	1.333	97.721	1.67E-12
70	364.551	48.73	332.42	626.044	1.33	1	2.99	2	95.94	116.03	45.998	1.181	97.42	5.57E-14
71	631.905	48.707001	440.937	805.478	2.06	14.31	143.48	4	182.91	194.005	76.91	1.14	133.363	
72			994.3	1720.5				18	289.45					
73	297.184	47.879002	257.368	501.946	3.68	401.39	2542.35	1	103.55	96.909	38.418	1.25	77.087	4.02E-09
74			33.9	116.1	0.45	1.29	41.84	0	12.53				33.98	2.20E+01

Chapter 5. Results of Database Analysis by Minitab using Multivariate analysis

5.1 Introduction

This chapter gives the results of the analysis performed by Minitab (version 16). Analysis shown in this chapter is dendrogram analysis and PCA for the three databases whose construction was described in chapter 4. Chapter 5 aims to answer the Research Question RQ2: What is meant by the term 'fundamental science' in relation to process plant cleaning? Chapter 4 gave an insight into what the term means, but the aim of this chapter is to fully answer the question. This will be carried out by examining the multivariate analysis from the dendrograms and PCA for the appropriate database of variables.

Section 5.2 of this chapter discusses initial results obtained from carrying out multivariate analysis using Minitab software and analysing data using dendrograms.

Section 5.3 examines and discusses the results from analysing the variables using PCA on the functional and structural properties of known API's. This is the information contained in database 1 (This information is listed in appendix II). In section 5.4 the results of the analysis by PCA on the second database containing information on the physicochemical properties of the same API's will be presented and discussed. In section 5.5 the PCA results from the third database containing the combined variables of databases one and two will be shown and discussed.

The results of both types of multivariate analysis used in this research will be further discussed in section 5.6. The choice of methodology to analyse the databases will be discussed and this will lead to the best choice of methodology and database to use for this research. This discussion will lead to a choice of database, which will be used to analyse industrial data provided by Britest members. Section 5.7 discusses the model development and begins to consider how cleaning agent and solvents data can be mapped onto the API data to create a more informative model, which will fulfil the remit of the aims of this thesis given in Chapter 1. In section 5.8 the purpose of the new model will be considered and how it can work with the existing remit of Britest tools discussed in Chapter 3. It is considered that the development of this model and its positioning in an adapted set of tools already discussed (chapter 3), will make a fundamental difference to the understanding of process plant

cleaning. Section 5.9 provides a summary of this chapter. A conclusion of this chapter is also provided in section 5.10.

5.2 Multivariate analysis – Initial results - Dendrograms

This method (as described in chapter 4) was used to analyse database 2 containing data on the physicochemical properties of the chosen API (listed in Appendix III). The horizontal axis of the dendrogram on figures 5-1 and 5-2 represent the distance between the clusters, or how much they are dissimilar. These figures show the amalgamation of information and suggest that clustering can be used to identify distinct groupings with similarity. One of the main discerning features of using this technique is that it can show discrete clusters. Figure 5-1 indicates before the final cut the clusters are more discrete but have a lower similarity. Figure 5-2 showing the final cut indicated that the clusters formed had a small number of variables.

This method was used as an initial analysis of the data to determine whether any patterns or similarities could be observed in the information. A dendrogram analysing by single linkage was used because it looks for the closest distance between points. Absolute correlation coefficient distance was chosen because it shows the relationship between variables in the data. Figure 5-1 shows the greater the similarity the stronger the relationship between the data.

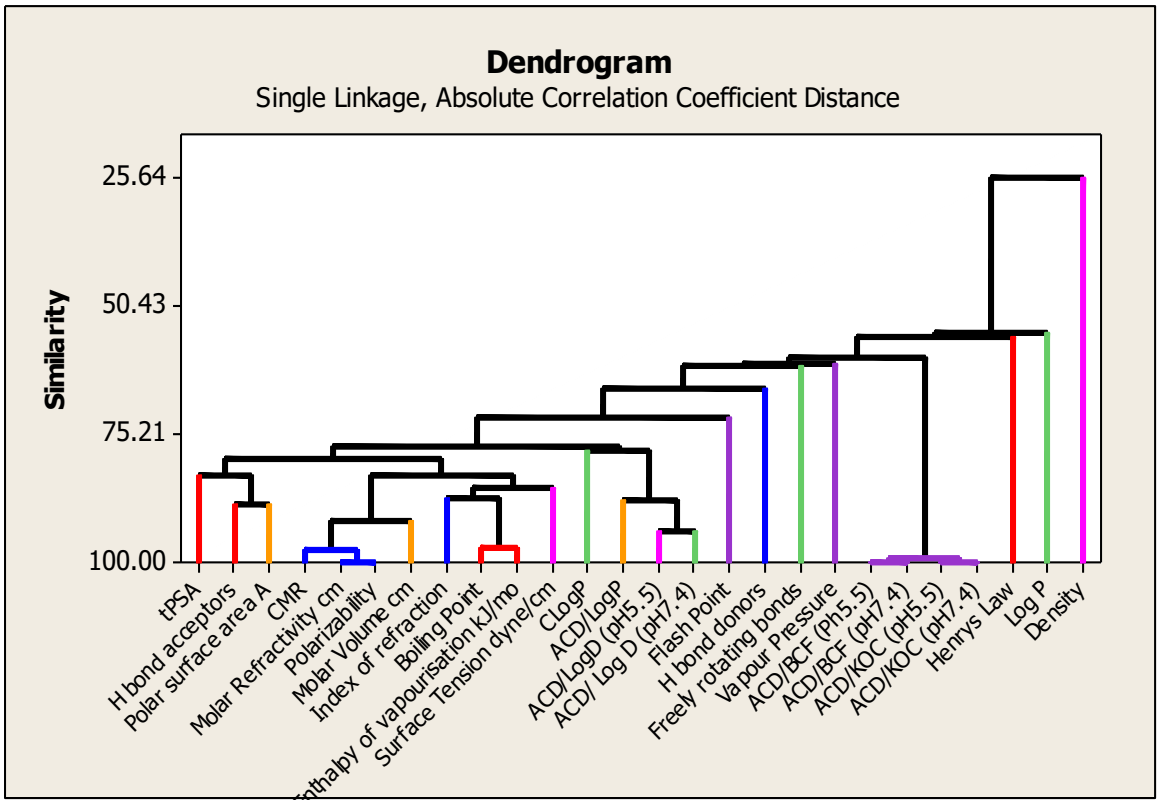


Figure 5-1 Dendrogram of data in database two relating to chemical properties of Britest member's pharmaceutical products and ingredients. The information is showing clustering of variables according to similarity prior to the final cut.

Figure 5-1 indicates that there was similarity in the data and a number of key clusters are shown. It is possible to determine that there is a similarity between data ACD/BCF (pH5.5), ACD/BCF (pH7.4), ACD/KOC (pH5.5) and ACD/KOC (pH7.4). This information can be described as follows - Advanced Chemistry Development Inc (ACD) is a company, which developed software for NMR prediction, nomenclature, chemical structure drawing, and physicochemical property prediction. The clustered information in figure 5-1 refers to tests at various stated pH values for the bioconcentration factor (BCF) and soil absorption coefficient (KOC). BCF - which is the term given to the concentration of a contaminant in or on a water organism such as fish. Bioaccumulation tests can use bioconcentration factors (BCF) to predict the concentrations of hydrophobic contaminants in organisms. BCF is the ratio of the average concentration of test chemical accumulated in the tissue of the test organism (under steady state conditions) to the average measured concentration in the water (Schäfer, 2015). A high BCF figure indicates low solubility of that particular chemical. KOC is the term indicating the tendency of a chemical to bind to, or adsorb to soil, per amount of water. Chemicals with large KOC figures tend to bind to soil (reach-serv, 2016). The clustering of this information is interesting, as it relates to the solubility of the chemical in water, which is

one of the factors that was discussed in chapter 3 as being important in terms of cleanability. Two other clusters of data, visible on figure 5-1, are molar refractivity and polarizability. These two variables are connected, so it is appropriate that they should form a cluster on the dendrogram. In order to explain this it is necessary to give a definition for molar refractivity. Molar refractivity is a measure of the total polarizability of a mole of a substance. Polarizability is '*a measure of the ease with which the electron distribution in a molecule can shift in response to a change in electric field; the ability of an atom to accommodate a change in electron density*' (Fox, 2016). The value for the variable molar refractivity takes into account the value of total polarizability. Therefore once this was known it was likely to assume that these variables should cluster together. Both of these variables were similar to CMR on figure 5-1, indicating they are similar. Another cluster identified on figure 5-1 was boiling point and enthalpy of vaporisation. Boiling point indicates the temperature at which a chemical boils. Enthalpy of vaporisation is the energy which needs to be expended to turn a liquid into a gas. There is a relationship between the energy which is needed to convert a liquid to a gas and the boiling point of a chemical, and therefore a close relationship would be expected between these two variables. In order to analyse the data further it was necessary to cut the data to give clusters shown in figure 5-2.

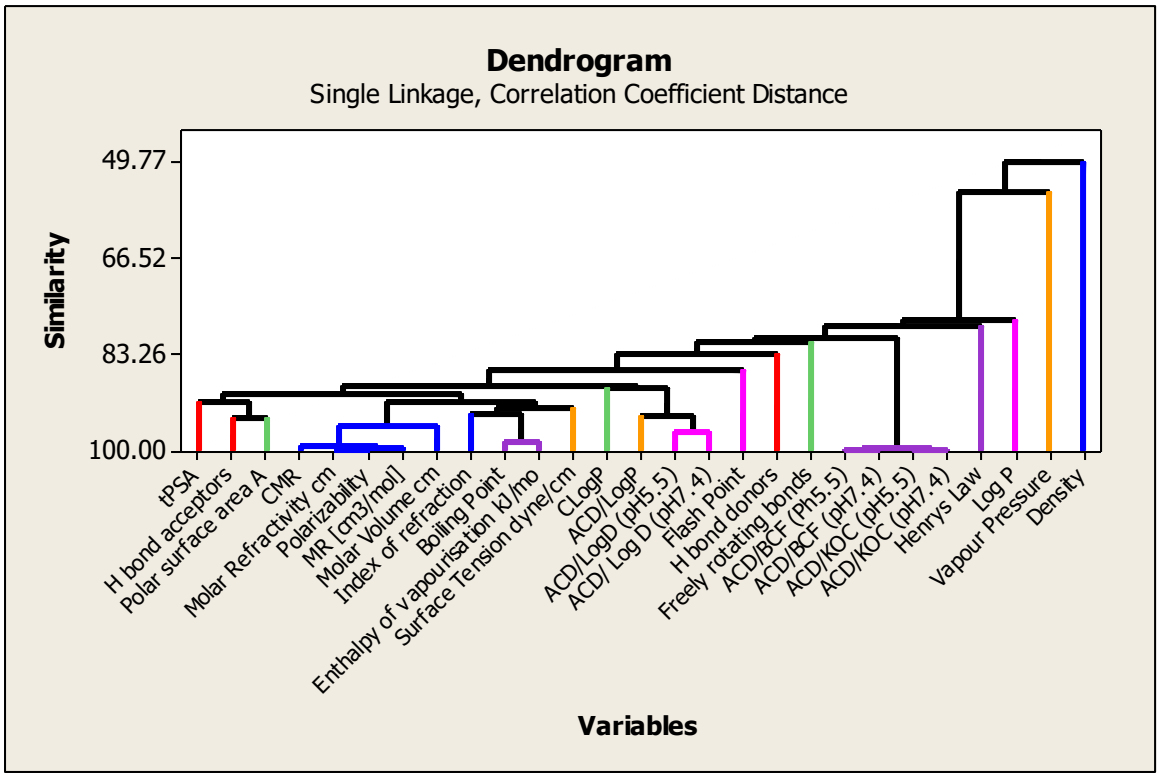


Figure 5-2 Dendrogram of data in database two relating to chemical properties of Britest member's pharmaceutical products and ingredients. The information showed clustering of variables according to similarity after the final cut.

The final cut taken for this specific set of data was indicated in figure 5-2. This figure shows a higher similarity between the variables. Clusters of interest were considered to be those with a similarity greater than 80%, as shown on figure 5-2. These clusters were ACD/KOC (pH7.4), ACD/KOC (pH5.5), ACD/BCF (pH5.5) and ACD/BCF (pH7.4), which was identified in the previous figure, and CMR, Molar Refractivity and Polarizability. In addition, in the same cluster, relative molecular mass (MR) was identified as being similar. It is not known why this variable clustered in this position. Other clusters of variables identified were ACD/LogD (pH5.5) and ACD/LogD (pH7.4). In order to understand why these variables have clustered with a high similarity, it is important to consider what the term LogD means. A definition is given by the ACD website '(*Log*) *D is the distribution coefficient and is a pH dependant measure of the propensity of a molecule to differentially dissolve in two immiscible phases, taking into account all ionized and unionized forms (micro species). It serves as a quantitative descriptor of lipophilicity' (ACD Inc, 2016).* LogD is a useful variable to know and understand in the pharmaceutical sector, because it can be used to assess drug likeness, and also in pharmacokinetics to help determine the ability of a drug to be absorbed, metabolised and also excreted. These variables relate to the solubility of a chemical, which is an important

consideration in plant cleaning as described in chapter 2. Closely related in similarity to the variable LogD is the variable LogP. LogP is described where P is the partition constant and is a measure of the propensity of a neutral molecule to differentially dissolve in two immiscible phases. It also serves as a quantitative descriptor of lipophilicity. This variable indicates the ability of a drug or chemical to be absorbed and it can also be used to assess drug likeness. This variable is also associated with solubility of chemicals.

Another interesting cluster of variables were Hydrogen (H) bond acceptors and polar surface area. These two variables are defined as follows - H bond acceptors are molecules with the ability to accept Hydrogen bonds, and polar surface area is the total surface area over all the polar atoms. Polar atoms include oxygen and nitrogen and their associated H atoms (Clayden et al, 2001). Associated with this cluster is the variable tPSA. Topological polar surface area (tPSA) is defined as a measure of polar surface area (Prasanna, 2009). It is therefore logical that the variable tPSA would be associated with the variable polar surface area.

Boiling point and enthalpy of vaporisation were variables associated in a cluster in figure 5-1. In addition to this figure 5-2 shows other variables associated with this cluster. These were index of refraction and surface tension. Index of refraction is the number that refers to the ability of light to travel through a medium. Surface Tension is the tension of the surface film of a liquid, which is caused by the attraction of the particles in the surface layer by the majority of the liquid. This tends to minimise the surface area (Clint, 1992). There seems to be no physical, chemical or other explanation for the similarity of these variables in the analysis.

In addition to the variables described other clusters are visible the hierarchy of clustering is visible on figure 5-2. These will not be described further.

The dendrogram cluster analysis has indicated that it may be a useful tool to cluster information when beginning to analyse cleaning methodologies. In order to determine if this is the best method to use for this analysis, it is important to consider other forms of analysis for the same data. In order to decide the best methodology for the examination of the data, a further methodology based on multivariate analysis was explored. Principal component analysis was investigated in section 5.3.

5.3 Principal Component Analysis Results and Discussion

5.3.1 Introduction

The method of principal component analysis (PCA) is discussed in chapter 4. In the next sections of chapter 5 the results of the analysis of each database 1 to 3 are presented and discussed. It is important to note for the purpose of clarity that some of the information and results are stated in this chapter as being of significance. The term significance within the body of this research means of greater impact and of greater weight than other results. This term was used with the understanding that within PCA analysis any results produced are evaluated by subjective decisions.

5.3.2 Introduction Database One Results and Analysis

Database 1 Results are shown in a series of figures (5-3–5-11). Where figure 5-3 shows a scree plot of the analysis, figure 5-4 shows the score plot of the analysis and figure 5-10 shows the loading plot of the analysis. It is considered important that each plot is examined in order to determine if the analysis of this data would make a good model or tool to aid industrialists in cleaning process equipment. The initial analysis which was examined was the scree plot in section 5.3.3.

5.3.3 Scree Plot examination for the PCA analysis carried out on Database 1 containing structural and functional group information on Active Pharmaceutical Ingredients.

The scree plot (figure 5-3) gives a visual plot of eigenvalues against principle component values. The scree plot was challenging to interpret but it was considered that the initial principle components of interest were in the initial 27 points. At 27 points the ‘elbow point’ of the plot is shown. The actual data which corresponds to the scree plot is shown in appendix V. Using this data it was possible to state that the first ten component numbers have an eigenvalue of greater than 2. It is often considered according to the Kaiser criterion that principal components with a value above 1 should be retained (Kaiser, 1960). In this research the principal components with a value greater than 1 were retained, but other criteria (it is common practice when using PCA to use several techniques to interpret the data (Joliffe, 2002)) were taken into account to decide which principal components to retain. This meant that some of the principal components not retained had values of greater than 1. This was due to the fact that the explained variance was a criterion used to determine the principal components.

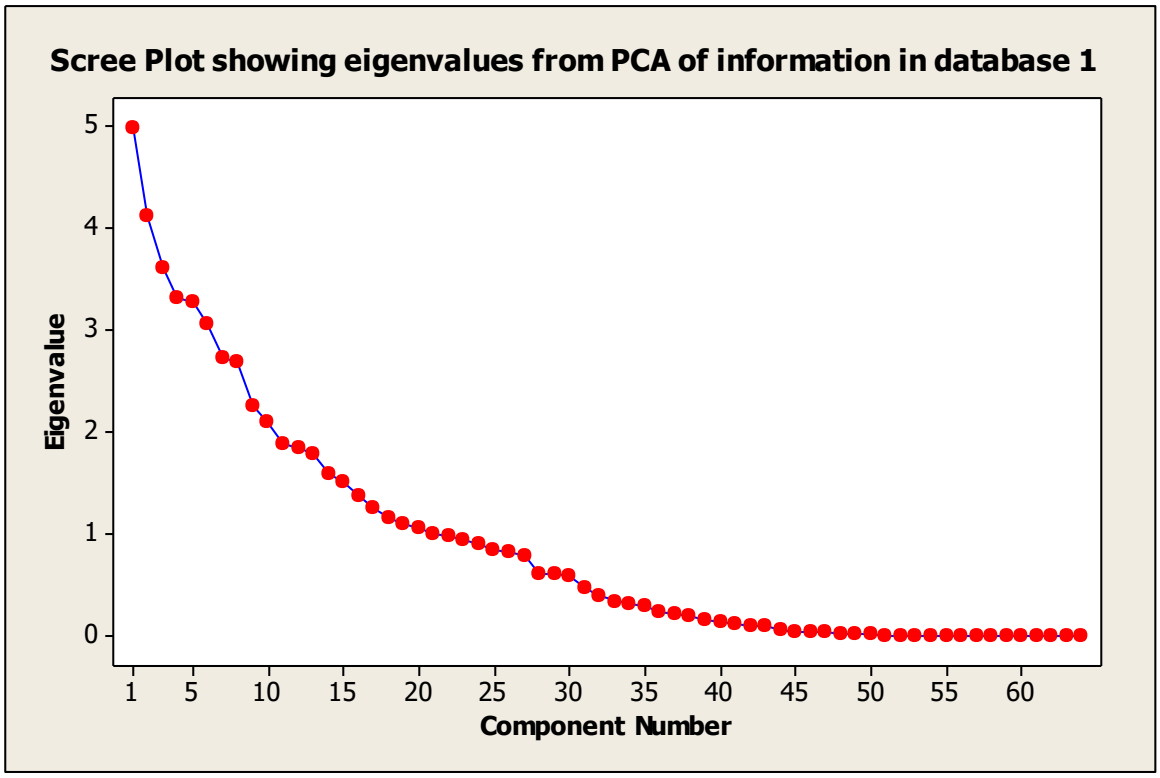


Figure 5-3 Scree plot from PCA of variables in database 1 on the functional groups and structural features of API's manufactured by Britest members.

The initial principle component produces the greatest total variation in the data set (eigenvalue 4.7752 and a total percent variation of 8.4%). The eigenvalues decrease in value after this point as would be expected, given that they account for less variation in the data. (The data used in the scree plot (figure 5-3) is shown in figure I in appendix V). The figure numbered I gave the eigenvalues for each principal component as shown in figure 5-3. It was necessary to consider that some of the principal components added less variation to the data. The components which explained 70% of the variance in the data were retained in this analysis. This accounted for the initial 14 principal components out of a total of 57 components examined. These principle components were taken to be of interest when examining the rest of the data and score and loading plot respectively for database 1. This was because the scree plot had determined the principal components and this strongly relates to the information in both the score and loading plot.

In addition to the criteria chosen to determine the principal components, eigenvalues for each principal component were considered to be insignificant if they had an eigenvalue of below 0.150. Eigenvalues above 0.150 were determined as significant and negative eigenvalues greater than -0.150 were also considered significant values. These values were chosen because they gave a range of data determining the extremes of the data showing the greatest

variance. Outside of these boundaries eigenvalues were not considered as important to the research. This method was used to determine the eigenvalues considered the most significance for the first 14 principal components. This was carried out by examining the eigenvalues for each principal component and determining if the value was within the remit required (above 0.150 or below -0.150) to be considered significant. This gave an indication of which eigenvalues were significant to which principal component and this information could then be related back to the variables. The principal component values all related back to variable data and this was determined in each case. This gave the information in table I in appendix V.

The scree plots (figure 5.1) were used to identify the functional group and structural features of interest, shown in Table I (appendix V). This was carried out by relating the functional group and structural information back to the eigenvalues for the principal components.These features provide the most variation in the data set according to the scree plot analysis of the first 14 principal components.

The information produced from analysis of the scree plot was correlated back to the API's in the analysis by determining which API's contained the variable of interest to see if any clusters or groupings of API could be determined. This produced a list of API which contained the structural features, or functional group, associated with variation in the data. This information is shown in Table II in appendix V.

The information in table II (appendix V) showed information of interest in the scree plot and related it to the API's used in the analysis. The pharmaceuticals identified as having chemical functional groups or structural features contributing to the variability in the data set are given in table 5-1.

Pharmaceutical product identified	**Significant functional groups or structural features identified**
Betamethasone disodium phosphate	Na+ Association, Hydrozone, Phosphate group, Phosphonate group, Tertiary alcohol association
Bambec	Secondary amine group, Phenyl ring
Blopress	Carboxylic acid group, Phenyl ring
Brofen	Carboxylic acid group, Phenyl ring
Citanest	Secondary amide group, phenyl ring
Clarithromycin	Macrolide, Tertiary alcohol structure

Pharmaceutical product identified	Significant functional groups or structural features identified
Deflox	Phenyl ring, Aromatic enamine group
Doxcycline monohydrate	Tertiary alcohol group, Vinyl alcohol group
Epival	Na+ Association, Carboxylic acid group
Folic acid	Primary amine group, Secondary amide group, Aromatic enamine group, Carboxylic acid group
Furosemide	Carboxylic acid group , Phenyl ring, Secondary amine group
Gadopentetate dimeglumine	(Gadolinium) Gd3+ Association, Carboxylic acid group , Secondary amine group
Gadopentetate monomeglumine	(Gadolinium) Gd3+ Association, Carboxylic acid group , Secondary amine group
Gopten	Carboxylic acid group, Secondary amine group, Phenyl ring
HPMPC	Phosphonate group, Aromatic enamine group
Hytrin	Aromatic enamine group, Phenyl ring
Invermectin	Tertiary alcohol group, Macrolide
Klacid	Tertiary alcohol group, Macrolide
Levothyroxine	Primary amine group , Carboxylic acid group, Phenyl ring
Lupron	Aromatic enamine group, Secondary amide group , Phenyl ring
Marcaine	Secondary amine group, phenyl ring, Secondary amide group
Metrolazole	Secondary amine group, Phenyl ring
Oxis	Secondary amine group , Secondary amide group, Phenyl ring
Plendil	Phenyl ring, Aromatic enamine group
Quinapril	Carboxylic acid group, Secondary amide group
Roxithromycin	Tertiary alcohol group, Macrolide, Oxime group
Salmeterol xinafoate	Secondary amine group, Carboxylic acid

Pharmaceutical product identified	Significant functional groups or structural features identified
	group
Sevelamer	Primary amine group, Secondary amine group, Tertiary alcohol group
Teveten	Carboxylic acid group , Hydrozone group
Warfarin	Vinyl alcohol group , Phenyl ring

Table 5-1 Pharmaceutical products, their associated chemical functional groups and structural features, which were identified as showing the most variation within the data set in database 1.

Table 5-1 shows the pharmaceutical products which were identified as having characteristics that generated the most variation within the data set. The most prominent functional groups in table 5-1 will be considered in this section. The characteristics identified in table 5-1 could include features of API's which could influence the cleanability of equipment. It is known that chemicals can be grouped according to chemical functional groups (Chapter 2 section 2.2.9) and this method shall be used to determine if there were any patterns or reasons why these characteristics could have been identified in this research. It was considered that certain functional groups contribute to water solubility and some of these functional groups are represented in table 5-1. These include hydroxyl or alcohol OH group and carbonyl groups (aldehyde groups and ketone groups). Table 5-1 shows that there were API's which contain different functional groups which include hydroxyl groups. These include tertiary alcohol association and vinyl alcohol groups in several API's listed (Betamethasone disodium phosphate, Doxcycline monohydrate, Invermectin, Klacid and Sevelamer). Other types of alcohol groups were not found in these API's. The carbonyl functional groups were represented by secondary amide groups in table 5-1. These were present in the API's Citanest, Folic acid, Marcaine and Quinapril. There were no other carbonyl groups represented in the data in table 5-1. This may have been expected as they increase polarity and reactivity of molecules and therefore increase solubility. Amines are known to be very soluble in water and therefore API's containing amines should be easy to clean from surfaces during plant cleaning. This can depend on the size of the molecule as the hydrocarbon chain gets longer the solubility of the molecule decreases (Clark, 2004). Table 5-1 shows that there were several types of amine identified. These included primary amine groups (in which only one of the H groups is replaced, which have a higher boiling point than secondary and tertiary

amines because they can form hydrogen bonds with each other as well as van der waals and dipole – dipole interactions (Clark, 2004)). Primary amines were present in Folic acid, Levothyroxine and Sevelamer. In secondary amine groups, two of the hydrogen in an ammonia groups are replaced by hydrocarbon groups, and this means that their boiling point is lower than primary amines (Clark 2004). In table 5-1 these were present in Bambec, Furosemide, Gadopentetate dimeglumine, Gadopentetate monomeglumine, Gopten, Marcaine, Metrolazole, Oxis, Salmeterol xinafoate and Sevelamer. There were tertiary amine functional groups represented in table 5-1 (Betamethasone disodium phosphate, Invermectin, Roxithromycin, Sevelamer, Clarithromycin, Doxycycline monohydrate and Klacid). In these groups all of the hydrogen in an ammonia molecule have been replaced by hydrocarbon groups (Clarke 2004). The other amine group represented in table 5-1 was the aromatic/enamine group. This functional group is an unsaturated compound. It is relatively reactive and it is nucleophilic. This means that they can be converted into aldehydes and ketones by acid catalysed hydrolysis (Clayden, 2001). This group was represented in API's HPMPC, Plendil, Deflox and Folic acid in table 5-1.

Other functional groups represented in table 5-1 included the acid groups, which were represented by acidic functional groups. The carboxylic acid group was the only group represented in table 5-1. Carboxylic acids are organic acids that contain a carbon atom that participate in both a hydroxyl and a carbonyl functional group. These functional groups can hydrogen bond with themselves in non-polar solvents, which raises the boiling points of API's they are connected to (Clayden, 2001). Therefore, it could be considered that these functional groups are influential in the structures of API's. This is because raising the boiling point of an API will change the ability to remove it from process equipment. Carboxylic acid functional groups were present in several API's including Gadopentetate monomeglumine, Gopten, Levothyroxine, Quinapril, Salmeterol xinafoate, Teveten, Blopress, Brofen, Epival, Folic acid, Furosemide and Gadopentetate dimeglumine.

In addition to the functional groups identified in table 5-1 there were several structural features. These were macrolides and phenyl rings and associations with Gd3+. Macrolides are a class of antibiotics which are bacteriostatic. They inhibit the growth of bacteria by inhibiting bacterial protein synthesis (Schlecht, 2016). Marcolides are large complex mixtures of closely related antibiotics and are basic in nature. They are poorly water soluble but they do dissolve in more polar organic solvents. Macrolides have a number of functional groups and this makes it possible for them to take part in multiple chemical reactions (MSD, 2015). The fact

that they are poorly soluble means that it is not surprising that they have been identified in the analysis, as this will have an effect on the ability to clean these API from vessels post manufacture. The API's which contain macrolides in this research are indicated in table 5-1, and include the antibiotics Clarithromycin, Invermectin, Klacid and Roxithromycin. There are several classes of antibiotics represented by API's used in this research, see Appendix II. Each antibiotic has different structures, properties and modes of action and it is possible that each class of antibiotic requires very different cleaning agents to remove it from vessels post manufacture.

Another structural feature which was identified in table 5-1 was the phenyl ring functional group. Phenyl rings are very common in the API's used in this research, which is not surprising as the formation of a phenyl ring gives a very stable structure which is required as a drug property. Phenyl rings are formed when a benzene ring is attached to a molecule by only one of its carbon atoms (Clayden, 2001). Phenyl rings are considered to be hydrophobic and they are therefore unlikely to be cleaned easily from production vessels using water alone. It is considered that the presence of a phenyl ring may have an influence on how an API may be removed from a production vessel by cleaning and choice of cleaning agent. The structural feature is found in many API's in this research, see Table 5-1, these include Warfarin, Plendil, Oxis, Metrolazole, Marcaine, Lupron, Levothroxine, Hytrin, Gopten, Furosemide, Deflox, Citanest, Brofen, Blopress and Bambec.

The final structural feature which will be discussed in relation to table 5-1 is Gadolinium (Gd^{3+}) association. This was identified in the pharmaceutical products Gadopentetate monomeglumine and Gadopentetate dimeglumine. Both of these API's are used in Magnetic Resonance Imaging (MRI) as contrast agents. Therefore they need to be soluble in the human body. These API's are described as freely soluble in water (O'Neil, 2013) and therefore the ability of cleaning API's associated with Gd^{3+} could be affected by this.

In Table 5-1 also pharmaceutical products (API's) can be found which have more than two features of interest. These products include Betamethasone disodium phosphate (Tertiary alcohol association, Hydrozone, Phosphonate group, Phosphate group and Na^{+} association), Folic acid (Primary amine group, secondary amide group, aromatic/ enamine group and carboxylic acid group), Gadopentetate dimeglumine (Gd^{3+} association, carboxylic acid group and secondary amine group), Gadopentetate monomeglumine (Gd^{3+} association, carboxylic acid group and secondary amine group), Gopten (carboxylic acid group, secondary amine group and phenyl ring), Levothyroxine (Primary amine group, carboxylic acid group and

phenyl ring), Lupron (Aromatic enamine group, Secondary amide group and Phenyl ring), Marcaine (Secondary amine group, phenyl ring and Secondary amide group), Oxis (Secondary amine group , Secondary amide group and Phenyl ring) and Sevelamer (Primary amine group, Secondary amine group, Tertiary alcohol group). It is not known which functional groups dominate the properties of the API over others and therefore could significantly influence the choice of cleaning method or agent.

A number of API's (23) have not been identified in the analysis and are not shown in table 5-1. These are comprised in table III in appendix V. The features of these APIs do not greatly contribute to variation within the dataset, although some of the products showed similar chemical functional groups. Commonly occurring functional groups in products included the functional groups esters, which were present in eleven products; ketone functional groups, which were present in eleven products; steroid features, which were found in ten products; secondary alcohol functional groups, which were present in ten products, and ether functional groupswhich were present in nine products among others listed in appendix V in table III.

The significance of the properties associated with the structural features and functional groups will be examined and discussed later in chapter 5. First, it is important to further analyse the information in database 1 created by PCA, by examining the score plot in section 5.3.4.

5.3.4 Score plot examination for the PCA analysis carried out on Database 1 containing structural and functional group information on Active Pharmaceutical Ingredients.

The score plot (figure 5-4), related to data associated with chemical functional groups in a series of API's or pharmaceutical products. The score plot describes the relationship between the data, in this case based on the relationships between the first and second principal components.

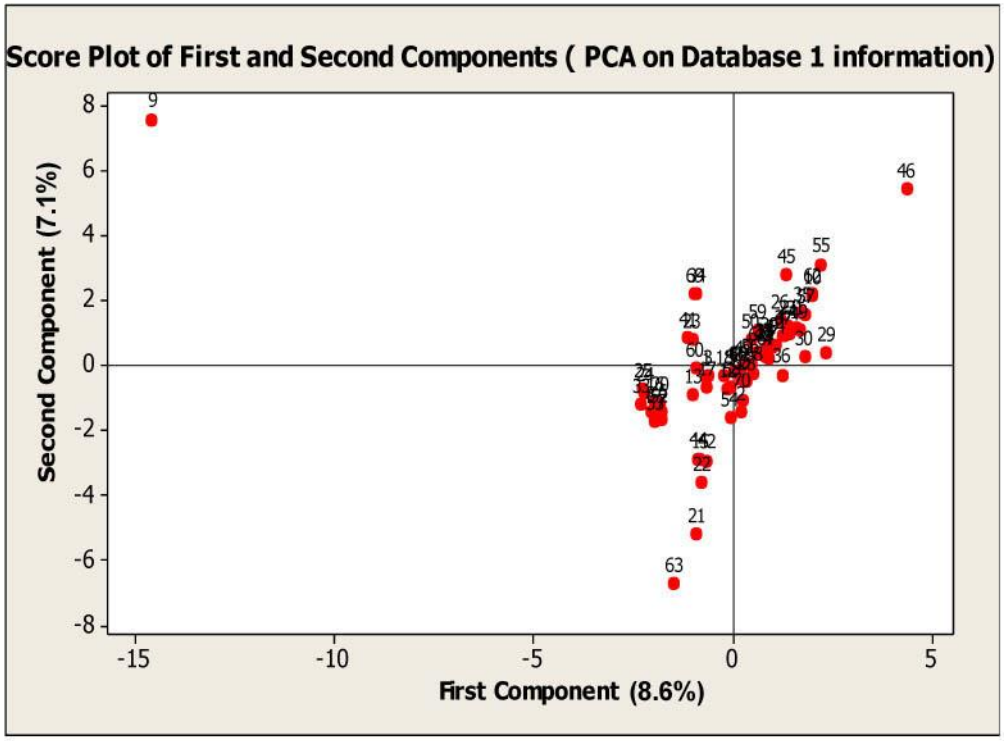


Figure 5-4 Score plot showing data associated with chemical functional groups in a series of pharmaceutical products manufactured by Britest members. The numbers shown on the plot are row numbers used in the analysis which relate to different API's. (Appendix V).

Figure 5-4 clearly shows that the data was clustered, linked and separated based on the relationships found within the data set based on the first two principal components. It is apparent that a large amount of API's were clustered around the zero point on both axes. In order to identify the API's in each cluster, the original row numbers given to the API's were used to identify them on the score plot. Figure 5-4 indicates that variation is found in the data, however the first and second principal components account for only 15.4% of the variation in the data set. It is clear that within the first two principal components there are groups of points that indicate clear separate distributions in the data. In order to discuss this data clearly it is important to reproduce the plot and indicate possible groupings (Figure 5-5)

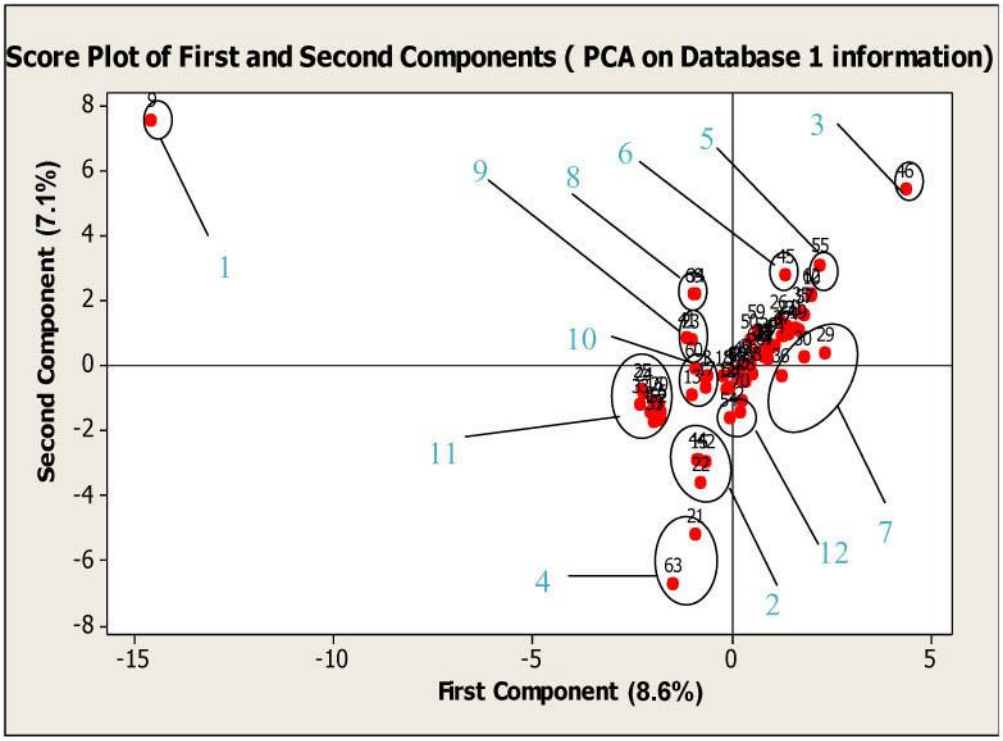


Figure 5-5 Score plot (figure 5-4) reproduced with annotation.

Figure 5-5 shows there were several identifiable groupings and prominent features which were circled. The circling or clustering was carried out by visual inspection which determined identifiable groups. A number given post the product refers to the pharmaceutical product reference number (in black font) the number given in blue font identified the group or cluster). These were identified in Table 5-2.

Identified Group or prominent feature	**Identified Pharmaceutical products**
1	Betamethasone disodium phosphate (9)
2	Clarithromycin (15), Invermectin (42), Doxcycline hyclate (22), Klacid (44)
3	Lupron (46)
4	Doxcycline hyclate (21)Roxithromycin(63)
5	Nizatidine (55)
6	Levothyroxine (45)
7	Gadopentetate dimeglumine (29), Gadopentetate monomeglumine (30) Imdur (36)
8	HPMPC (34),Teveten (69)
9	Epival (23), Isradipine (41)

Identified Group or prominent feature	Identified Pharmaceutical products
10	Advicor (1), Androgel (3), Ciclesonide (13), Conholip (17), Progesterone (60)
11	Beclomethasone dipropionate (6), Beclomethasone dipropionate monohydrate (7), Betamethasone acetate (8), Clobetasol propionate (16), Dexamethasone dipropionate (20), Fluticasone furoate (24), Fluticasone propionate (25), Halobetasol (33), Mometasone furoate anhydrous (52), Mometasone furoate monohydrate (53)
12	Aluvia (2), Nimbex (54), Venlafaxine (70)
Main data set	Atenolol (4), Bambec (5), Blopress (10), Brofen (11), Calcijex (12), Citanest (14), Cycloserine (18), Deflox (19), Folic acid (26), Furosemide (27), Gabapentin (28), Ciclosporin (31), Hytrin (35), Iodixanol (37), Iopanidol (39), Isoflurane (40), Marcaine (47), Meperidine (48), Meprobamate (49), Methohexital (50), Olanzapine (56), Oxis (57), Paricalcitol (58), Plendil (59), Quinapril (61), Ranitidine (62), Salmeterol xinafoate (64), Severane (66), Tamsulosin (68), Warfarin (71)
Products not identified in analysis	Sumatriptan Base (67), Selelamer (65), Gopten (32), Iohexol (38), Metronazole (51)

Table 5-2 Identified clusters and prominent features within score plot.

There are a few points that stand alone as statistically relevant points of interest and these are Betamethasone disodium phosphate (nine) and Lupron (46). It is considered that there is something considerably different about these API's which may mean they are cleaned from vessels differently from other API's identified in the score plot. This may be due to the contribution of functional groups and structural features in each API (given in table IV in appendix V). There are also two products which are outside of the main group, these are groups five (Nizatidine (55)) and six (Levothyroxine (45)). A series of small clusters has been identified. These are shown in table IV (in appendix V) as groups seven, eight, nine, ten, and twelve. Two larger clusters of data are shown in the score plot (figure 5-5) and these are labelled as group eleven and the main data set. The main data set clusters close to and around

zero. These comprise the majority of products. Several pharmaceutical products do not appear in the data set. These are also indicated in table 5-2.

Groups of products were identified in the score plot and this information was back related to the pharmaceutical products and the chemical functional groups and structures they contain, see Table IV in appendix V. Some of the characteristics in this table are primary characteristics, which were identified from the scree plot (table 5-1). In addition to these characteristics, a further set of characteristics was identified from the score plot analysis. Table IV, appendix V shows features that were determined in these groups from the score plot. These shall be called the secondary characteristics. In addition to the primary characteristics determined by the scree plot, the secondary characteristics will be discussed in the identified groups. This will help to determine whether these secondary characteristics are of importance in the score plot. The functional and structural information of the pharmaceutical products in identified groups on the score plot shall be analysed as follows.

Group 1

Group 1 contained one point of interest relating to the pharmaceutical product Betamethasone disodium phosphate (Figure 5-5). This is significantly different to the other data on the score plot. The first component is -14.5365 and the second component is -7.54679 and it lies in the upper left quadrant of the score plot. No other pharmaceutical product scores as low as this value. This product contains a significant number of groups identified as of interest in the scree plot (table IV, appendix V). It can be stated that the variance showed by this point is greater than the other points due to this factor. The groups of interest in this pharmaceutical product exceed any other group. These groups or features of interest are Na+ association, (it is associated with two Na+, unlike the product Epival in group nine which is associated with one), and it also has a hydrozone feature and tertiary alcohol association. In addition the product is associated with both phosphate and phosphonate groups. These features make it a unique point within the data set. Secondary characteristics identified in this group include secondary alcohol, ketone, Aryl halide functional groups and steroid structures.

Group 2

Group 2 consists of four different pharmaceutical products which lie in the bottom left quadrant of the score plot (figure 5-5). These data points refer to pharmaceutical products Clarithromycin, Invermectin and Doxcycline monohydrate and Klacid. These four products all have Tertiary alcohol functional groups within their structures that give them a common

link. The other feature associated with Clarithromycin, Invermectin and Klacid is a structural feature macrolide. The two pharmaceutical products Klacid and Clarithromycin have the same structural and functional group properties as they are the same product. This data was included in the dataset as an internal control. The pharmaceutical product Doxycycline monohydrate does not have this feature, although Doxycycline monohydrate has the presence of vinyl alcohol. In terms of secondary features in Group Two there are a mix of different functional groups and structural features. The group of products as a whole all contain secondary alcohol functional groups and three of the set contain both Esters and Ether groups. This set of data does contain products which have large numbers of Ether functional groups in comparison to the other pharmaceutical products. Clarithromycin and Klacid have six Ether functional groups each and Invermectin has nine.

Group 3

Group 3 consists of one pharmaceutical product which is Lupron. This product lies distinct in the right hand quadrant of the score plot (Figure 5-5). Its co-ordinates are given as first component 4.4029 and its second component is 5.41164. This product adds significantly to the variation within the dataset. This is composed of several features which were identified on the scree plot as being of significance. These features include aromatic enamine, secondary amide and a phenyl ring. It is an interesting product because it has multiple features in combination that makes it different to other API's in the data set. In terms of secondary features this product contains primary alcohol groups, phenol groups, secondary amide, guanidine, alkyl groups greater than 5 carbons and N-heterocyclic features.

Group 4

Group 4 consists of two pharmaceutical products, which are Doxcycline hyclate and Roxithromycin. This group lies in the bottom left quadrille of the score plot. It is quite close to the position of Group 2. The pharmaceutical products in this group, similar to Group 2 also have tertiary alcohol structures in their construct. Both products have additional groups identified by the scree plot as being of interest. Doxcycline hydrate has an associated vinyl alcohol feature, which makes it similar to Doxcycline monohydrate. These products are close in position on the score plot. Roxithromycin is a product, which contains an oxime group. It is the only product within the dataset to contain this feature, which means that the oxime group could have an effect on variation within the dataset. This is because it is the oxime group which makes it different from the other API's in the data set. The oxime group is known to be

very poorly soluble in water (Clayden, 2001) which may account for the difference in terms of solubility. This may then affect the ability to clean this product from equipment post manufacturing. Group 4 has two pharmaceutical products with very different functional groups in the secondary characteristics. Both products (Doxcycline hyclate and Roxithromycin) have tertiary amine functional groups.

Group 5

Group 5 consists of one pharmaceutical product, and although it lies close to the main central dataset, it is visually distinct from the main group. This product is Nizatidine. The only group of interest identified by the scree plot in this product is an aromatic/enamine. In terms of secondary characteristics functional groups include tertiary amines, thioesters, nitro groups. Secondary structural characteristics include N-heterocyclic and S-heterocyclic structures.

Group 6

Group 6 consists of a single pharmaceutical product closely associated on the score plot to Nizatidine, which is in Group Five. This product does not contain features identified by the scree plot as being of interest. The features associated with this product Levothyroxine are a primary amine, carboxyl acid and a phenyl ring. This product is different from other API's in the data set as it is a hormone with multiple functional groups. The other hormone in the data set (Progesterone) has different functional groups to Levothyroxine and was found in Group 10. Secondary structural features include Phenol, Ether, Aryl halide groups and the structural feature of a hormone.

Group 7

Group 7 consists of three pharmaceutical products. These are Gadopentetate dimegumine, Gadopentate monomeglumine and Imdur. The first two products are very similar in construction. Both products contain secondary amine, carboxylic acid, GD3+ association and secondary amides. In addition to this Gadopentate monomeglumine contains water. The third product in this group, Imdur, is interesting as it contains no features identified by the scree plot analysis as adding significantly to this variability within the data. It has a variety of features, which include secondary alcohol, ether, nitrate and O- heterocyclic. This data point does lie significantly close to the main data set in comparison to the other two products. However, given the absence of common features it could be stated that this point belongs in its own category. The only common feature in the secondary characteristics of interest is Secondary alcohol functional groups.

Group 8

Group 8 contained two pharmaceutical products, which are Hydroxyl 2 Phosphonomethoxy Propyl Cytosine (HPMPC) and Teveten. Both these products contain different features identified by the scree plot as contributing a high degree of variation to the dataset. HPMPC contains phosphonate and aromatic groups. Teveten contains carboxylic acid and hydrozone features. The two products lie very close to each other on the score plot. The reason for this will be investigated at a later stage in this report. The primary and secondary characteristics of these two products are not similar as indicated in table IV, appendix V.

Group 9

Group 9 contains the pharmaceutical products Epival and Isradipine. Epival has features which have been identified by the scree plot analysis as adding to the variation, these include carboxylic acid and Na+ associations. Isradipine does not contain any features identified during the scree plot analysis. It has features which include ester, pyridine, alkyl >5 carbons and N-heterocyclic structures. Therefore it is not known why this product clustered within this group. The primary and secondary characteristics of these two products are not similar as indicated in table IV, appendix V.

Group 10

The pharmaceutical products identified within Group 10 are Advicor, Androgel, Ciclesonide, Conholip and Progesterone. Table 5-2 shows none of these products are associated with the characteristics that have been identified as adding to the variability as described by the scree plot. There are no common primary or secondary characteristics. In addition, it can be stated that both Advicor and Ciclesonide contain Ethers, Ciclesonide and Progesterone both contain Esters, Ciclesonide and Androgel both contain secondary alcohols and Progesterone, Conholip and Ciclesonide all contain Ketone groups. Therefore the reason why it should cluster in this group was not determined at this point.

Group 11

Group 11 is a large group of pharmaceutical products which appear close to the main product group. With the exception of Betamethasone acetate, which contains a tertiary alcohol group, none of the identified products have any feature identified as significantly adding to the variation within the dataset. The other identified pharmaceutical products are Beclomethasone dipropionate, Beclomethasone dipropionate monohydrate, Clobetasol propionate,

Dexamethasone dipropionate, Fluticasone furoate, Fluticasone propionate, Halobetasol and Mometasone furoate monohydrate. The significance of the inclusion of Betamethasone acetate in this group will be discussed later in this report. Common secondary characteristics are found in all products, which are secondary alcohol groups, ketone, ester groups and steroid structural features. It may be that the accumulation of these secondary features is the reason for the clustering effect.

Group 12

Group 12 contains three pharmaceutical products. These are Aluvia, Nimbex and Venlafaxine. In this group there is only one product which has been identified by scree plot analysis as adding significantly to the variation in the dataset. This product is Aluvia, which contains primary amine features. Nimbex structure includes ester, ether, sulfone and N-heterocyclic features. One common secondary characteristic in all of these products is not found. However, both Nimbex and Aluvia contain Ester groups and both Venlafaxine and Nimbex contain Ether functional groups.

Main Dataset

The main dataset is located around the central point of the plot at the zero position. Table 5-2 shows it includes a lot of data points equating to a significant number of pharmaceutical products. These products are listed below with any common identified chemical functional groups or structural features given in brackets -

Atenolol, Bambec, Citanest, Mepridine (Phenyl ring)

Blopress and Brofen (Carboxylic acid, Phenyl ring)

Deflox, Plendil and Hytrin (Phenyl ring, Aromatic enamine)

Cycloserine (Primary amine)

Calcijex and Paricalcitol (Tertiary alcohol)

Marcaine and Oxis (Secondary amine, Secondary amide, Phenyl ring)

Folic acid, Salmeterol xinafoate and Furosemide (Secondary amine, Carboxylic acid)

Iodixanol, Iopanidol and Ciclosporin (Secondary amide)

Gabapentin (Primary amine, Secondary amide)

Quinapril (Carboxylic acid, Secondary amide)

Tamsulosin (Secondary amine)

Rantidine (Aromatic enamine)

Warfarin (Vinyl alcohol, Phenyl ring)

Methohexital, Meprobamate, Olanzapine, Isoflurane and Severane (No significant functional group or structural features identified by scree plot analysis).

In the main data set there is a wide range of secondary characteristics. No particular pattern of information is identifiable. Further analysis of the main group of products identified in section 5.3.4 was carried out and was discussed in section 5.3.5.

5.3.5 PCA of the main group of identified products

PCA was carried out using only the products in the main data set identified in section 5.3.4. This indicated the following results. The scree plot showed an elbow point after the first three principal components, all of which had eigenvalues greater than 2 (figure II, appendix V). Figure II indicated that this accounted for 33% of the variation in the data set. The first three variables in the dataset which account for this variability were the products Atenolol (15% of the variability in the data set), Bambec (10% of the variation in the data set) and Citanest (7.8% of the variation in the data set). All of these variables had phenyl rings identified in their structures. Other variables accounting for the variability in the data set with eigenvalues over 1 (taking into account Kaisers criterion), include Meperidine (also with a phenyl ring in its structure) (eigenvalue of 2.1344), Blopress (eigenvalue of 1.9974), Brofen (eigenvalue of 1.8286), Deflox (eigenvalue of 1.6475), Plendil (eigenvalue of 1.4493), Hytrin (eigenvalue of 1.3736), Cycloserine (eigenvalue of 1.2332), Calcijex (eigenvalue of 1.1429) and Paricalcitol (eigenvalue of 1.0156). Together these variables account for 79.4% of variation in the data set. The score plot of the same data (figure III, appendix V) shows the presence of individual and clustered data. It is possible to determine from figure III that a lot of the data is located around the zero point, which would indicate that a lot of the data shows no variation. There are points of data such as variable 2 (Bambec), variable 50 (Ciclosporin, the only macrocyclic product in the data set with a high number of secondary (4) and tertiary amides (7) and variable 16 (Salmeterol xinofoate (2 phenol alcohol groups, 1 long alkyl functional group and 1 ether group, 1 carboxylic acid group, 1 secondary alcohol group and 1 secondary amine group)), which are separate from the main data set. This indicates that there is variability in

the data associated for these variables. Analysis of the data suggests that the presence of one or more phenyl rings seems to account for a lot of the variation shown in the data set. The products Bambec and Atenolol both have one phenyl ring in their structures. In addition other similarities include the presence of secondary amine functional groups and secondary alcohol groups. The products are very different in some ways as Bambec has two carbamate functional groups and Atenolol has both primary amide and ether functional groups. Citanest also has 2 phenyl ring structure, and similarly to Bambec and Atenolol it has a secondary amine group. There is also a secondary amide functional group present. Meperidine has a phenyl ring and N-heterocyclic, tertiary amine and ester functional groups. Blopress is an interesting product within the data set as it contains 3 phenyl rings (and also 2 N-heterocyclic structures and an ether and a carboxylic group). It is possible to determine from this that the presence of a phenyl group may be of more significance than other variables in this data set. Analysis of the score plot (figure III, appendix V) shows that variables Atenolol (1) and Tamsulosin (53) have clustered closely together. The only similarity in the data is both have a secondary amine functional group. The variables Brofen (6) and Ipomidol (49) have clustered together, although there are no common functional groups between them. The Loading plot (figure IV, appendix V) showed the position of the products in relation to each other. The figure IV shows that the majority of the products which accounted for the highest variation in the data set are on the right hand side of the figure. The similarity of these products was discussed above and most contain a phenyl ring structure. The Loading plot shows Warfarin as one of the products on the right of the figure. This is possibly due to the fact that its structure contains 2 phenyl rings, but this product was not identified in the scree plot as being of high variation within the data set. It has other structural features such as an O-heterocyclic structure which contribute to the fact that it is not soluble in water (Melnikov, 1971), which may be the reason for this. It may be gathered from this data that although the main data set examined from figure 5-5 and table 5-2 was not considered of high relevance to the research in the first PCA, separate analysis shows that some variables (including phenyl ring structures and the presence of secondary amine functional groups are of relevance. It is important to determine whether analysis of the principal components by examination of the score plots PC3 and PC4 and PC5 and PC6 gives any more information on the variance within the data set. This will be discussed in section 5.3.6.

5.3.6 Analysis of further principal component score plots (PC3 v PC4 and PC5 v PC6).

In addition to the analysis carried out in section 5.3.5 it was important to consider the variance indicated in the other principal components to contribute significantly to the variation in the

data set. In order to achieve this, principal components 3 versus 4 (figure 5-6) and principal components 5 versus 6 (figure 5-8) were plotted. Principal components 1, 2, 3 and 4 combined together account for 27% of variation in the data. If principal components 1 to 6 are accumulated together this accounts for 40% of the variation occurring in the data set. The first plot under consideration is the score plot or scatter graph of the 3rd and 4th principal components (figure 5-6).

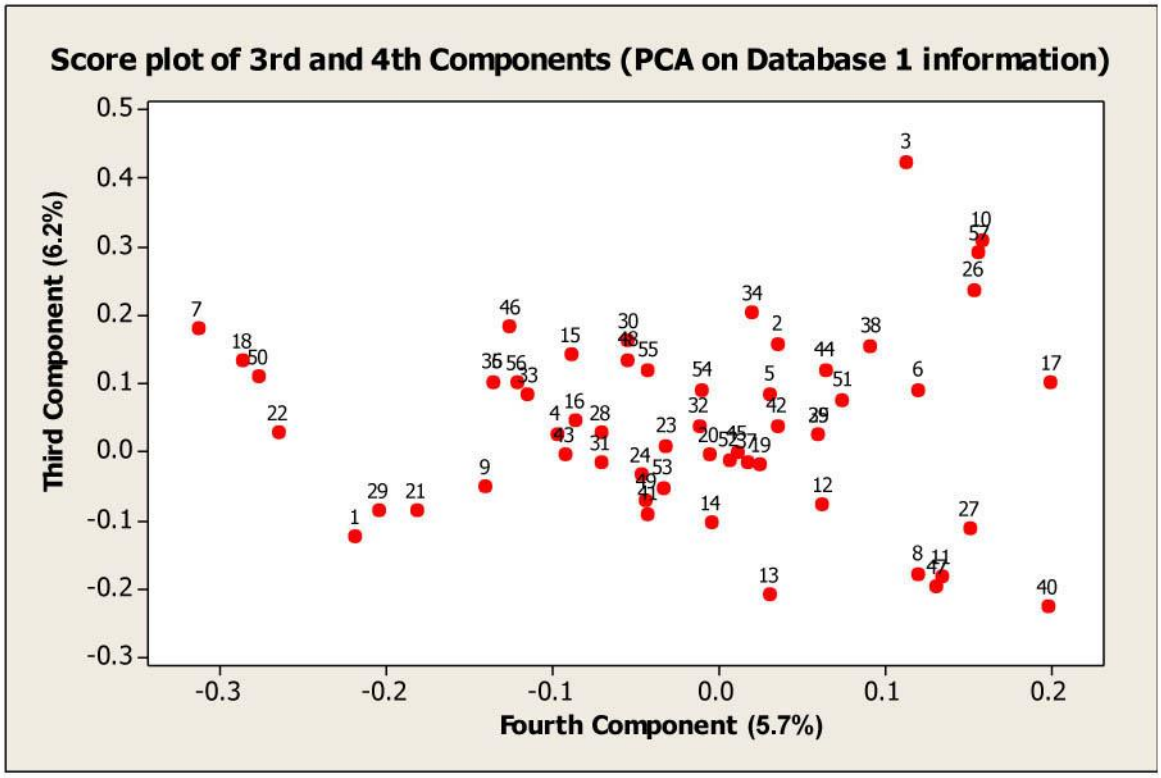


Figure 5-6 Score plot of Third and Fourth Principal Components showing data associated with chemical functional groups in a series of pharmaceutical products manufactured by Britest members. The numbers shown on the plot are row numbers used in the analysis which relate to different chemical functional groups and features. (Appendix V).

Figure 5.6 indicates some positioning of data around the zero point of both axes. It is possible to determine a number of clusters and groups within the dataset relating to the variables. Clustering is indicated by the addition of annotation (figure 5-7) and this is considered in table V, appendix V.

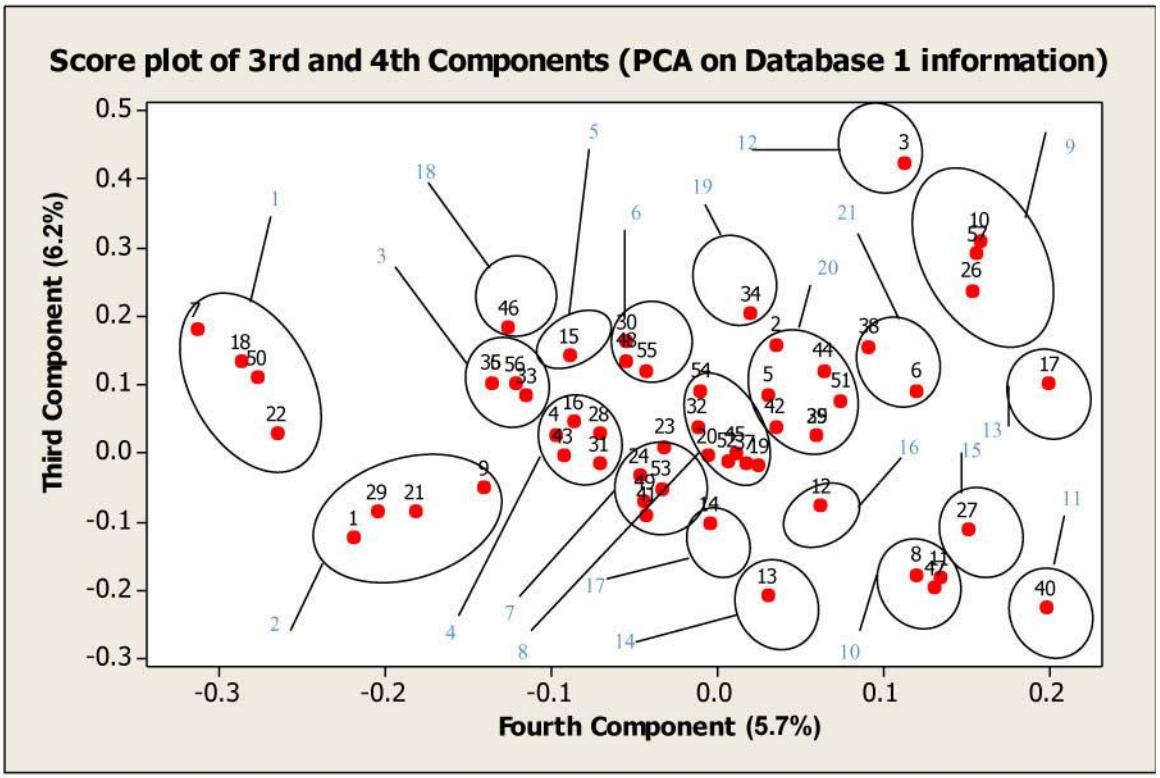


Figure 5-7 Score plot of the third and fourth components visualising groups and clusters identified by the analysis. Some of the groups of variables are identified in table V, appendix V.

Figure 5-7 and table V (appendix V) indicate that although it was possible to identify clusters in the data set, it was difficult to analyse the information. The distribution of the variables in figure 5-7 is broad. The variables which do appear to be clustered (table V, appendix V) correlated to information on variables found in API's in this research. It is possible to determine from table V (appendix V) that several API's have chemical functional groups or structural features present in more than one group. It is therefore not possible to give a similar interpretation of the results as for the score plot (figure 5-6). It was possible to determine from tableV (appendix V) that there are several distinct variables in the dataset. (For ease of clarity during analysis these shall be referred to as groups). These are considered to add significantly to the variability of the data in the principal components 3 and 4. These variables are listed in table V (appendix V) as group 5 (Secondary amide groups), group 11 (Steroid structure), group 12 (Tertiary amine groups), group 13 (Thioester groups) group 14 (Ester groups), group 15 (Fluorine groups), group 16 (Enone groups), group 17 (Primary amide groups), group 18 (Phenyl ring structure) and group 19 (Hydrozone structures). In addition the groups 1, 2, 9 and 10 and the variables they contain are of interest (table V, appendix V). In these groups it is possible to determine that there are several variables which correspond to the same API. In

group 1 Roxithromycin has all of the variables contained in the cluster (tertiary alcohol group, an oxime group, an ether group and macrolide structure). In addition this API has several other variables which have been identified in other groups within figure 5-7 and table V (appendix V). In group 2 (table V, appendix V). Lupron is identified more than once as having features and variables which are present in this cluster as it contains a guanidine group and a phenol group. Group 3 contains several features associated with Betamethasone disodium phosphate. These are a phosphate group, a phosphonate group and a Na+ group. It is possible to state that a majority of API's are associated with more than one group or cluster identified in figure 5-7 and table V (appendix V). After analysing figure 5-7 it is important to consider what a plot of principal components 5 vs principal components 6 can add to the analysis of the variables within the dataset. Principal components 5 and 6 were plotted (figure 5-8). This gave the following results.

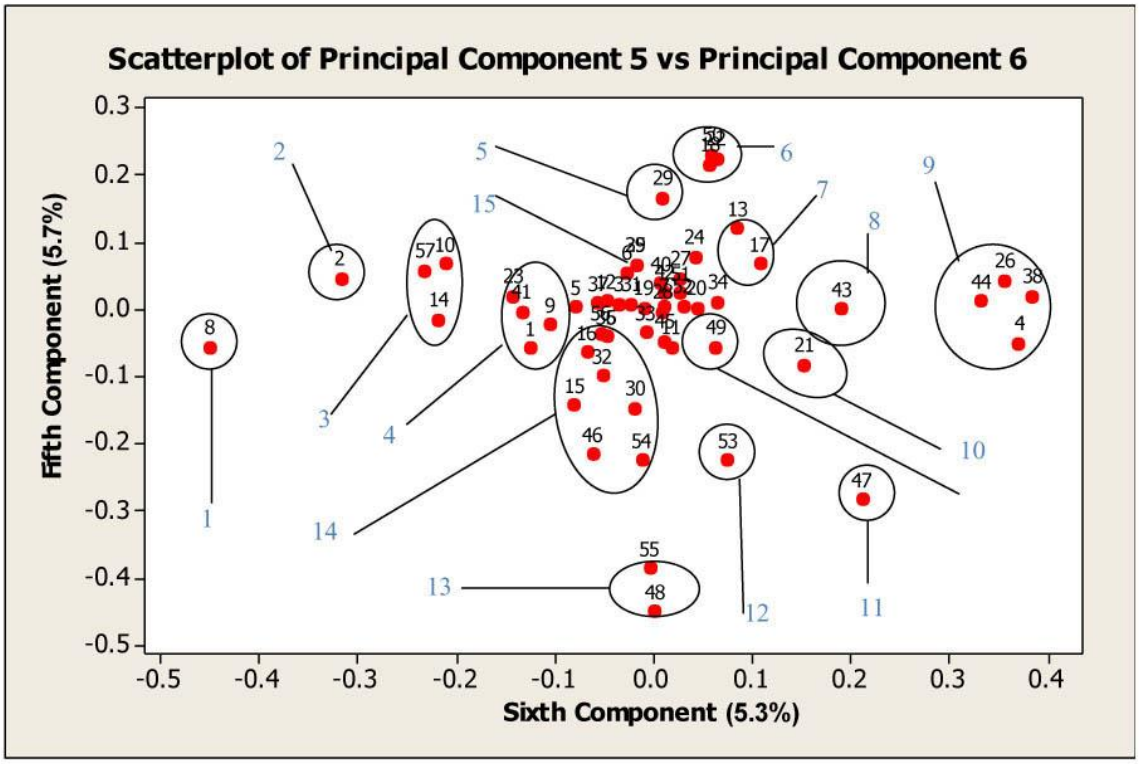


Figure 5-8 Score plot of Fifth and Sixth Principal Components showing data associated with chemical functional groups in a series of pharmaceutical products manufactured by Britest members. The numbers shown on the plot are row numbers used in the analysis which relate to different chemical functional groups and features. (Appendix V).

In order to analyse figure 5-8 it was necessary to annotate it to show potential clusters and groupings indicated within the data set by plotting principal components 5 versus principal components 6. Figure 5-8 indicates a number of clusters in the data which have been

identified by annotation. These clusters are listed in table VI, appendix V. Table VI shows the APIs associated with the variables associated with each of the identified clusters in figure 5-8.

There are a number of groups which could be associated with the variability in the dataset (table VI, appendix V). The table refers to the variability of the data in principal components 5 versus 6. Table VI (appendix V) indicates which variables add to the variability of the data and which of the API's in the dataset contain the variables. Groups of interest are the ones which are considered to be furthest away from the zero on both axes. These are considered to be group 1 (vinyl alcohol groups), group 2 (Secondary amine groups), group 3 (Gd3+, Carboxylic acid groups, Primary amide groups), group 8 (N-heterocyclic structures), group 9 (Aromatic/enamine, groups, Thioether groups, S-heterocyclic structural features, nitro groups), group 10 (Guanidine groups), group 11 (Erythromycin derivative structural features), group 12 (Water associated API's) and group13 (HCL and Tetracycline associated structural features). Several groups contain more than one variable associated with one API. These are group 3, which is associated with Gadopentetate monomeglumine and some of its variables Gd3+ structures, carboxylic acid groups and the variable primary amides. Group 9, which can be associated with Nizatidine for all variables which are contained in the group (Aromatic/enamine groups, Thioether groups, S-heterocyclic structural features and Guanidine groups). There are other variables associated with the API Nizatidine which are not associated with this group. These are nitro groups and N- heterocyclic structural features.

There are some variables which have been identified in other groups, which feature in the same API (Table VI, appendix V). An example of this is group 6 where the API Roxithromycin appears to contain all three variables associated with the group (Secondary amide, Ether group, Oxime group). Therefore, it is possible to state that the API's associated with these variables are present in more than one group as associated with the data in figure 5-7 and table V (appendix V). This information concerns the data plotted for principal component 3 versus principal component 4, or the information obtained from the initial figure 5-2, the score plot of principal component 1 versus principal component 2.

The individual data for each figure 5-2, 5-7 and 5-8 is interesting and using this, it was possible to suggest variables which may be considered to contribute to the most variation within the data set. This is considered in section 5.3.7.

5.3.7 Analysis of the first six principal components for database 1

This section aims to identify the variables within the first 6 principal components which add most to the variation within the data set. In order to analyse this information it important to identify the variables which were identified as adding considerably to the variability in each of the figures analysed in sections 5.3.4 and section 5.3.6. These were figure 5-2, PC1 versus PC2, figure 5-7, PC3 versus PC4 and figure 5-8 PC5 versus PC6. This is given in table 5-3 to 5-5.

Table 5-3 showing characteristics determined as important in adding to the variability of the data set according to PC1 versus PC2.

Variables					
Amine	**Alcohol OH**	**Acid**	**Carbonyl Groups**	**N Groups**	**Other characteristics**
Primary amine Secondary amine Aromatic/ enamine	Tertiary alcohol structure Vinyl alcohol	Carboxylic acid	Secondary amide	Oxime	Phosphonate Phosphate

Table 5-3 Identified functional group and structural variables within the PC1 and PC2 score plot analysis (figure 5-2).

Variables					
Amine	**Alcohol OH**	**Acid**	**Carbonyl groups**	**N groups**	**Other characteristics**
Tertiary amine group	Vinyl alcohol group	Carboxylic acid group	Ester group Thioester group Secondary amide group Enone group Primary amide group Ketone group	None identified	Steroid Fluorine group Phenyl ring Hydrozone structural feature Gd3+ group Thioether group Erythromycin derivative

Table 5-4 showing variables adding to the variability of the data in the principal components PC3 and PC4 score plot analysis (figure 5-7).

Variables					
Amine	**Alcohol OH**	**Acid**	**Carbonyl Groups**	**N Groups**	**Other characteristics**
Secondary amine group Aromatic/ Enamine groups	Non identified	Carboxylic acid group	Primary amide	Guanidine	Gd3+ group N-heterocyclic structures Thioether group S-heterocyclic structural features Water association Nitro group Erythromycin derivative Tetracycline structural features HCL association

Table 5-5 showing variables adding to the variability of the data in the principal components PC5 and PC6 (figure 5-9)

The information in tables 5-3 to 5-5 can be combined to give a list of variables of interest as determined by the first 6 principal components. The variables of interest are given in table 5-6.

Variables					
Amine	Alcohol OH	Acid	Carbonyl groups	N groups	Other characteristics
Primary amine Secondary amine Aromatic/ Enamine Tertiary amine group Secondary amine group	Vinyl alcohol group Tertiary alcohol structure	Carboxylic acid group	Primary amide Ester group Thioester group Secondary amide group Enone group Ketone group	Guanidine Oxime	Gd3+ group N-heterocyclic structures Thioether group S-heterocyclic structural features Water association Nitro group Erythromycin derivative Tetracycline structural features HCL association Steroid Fluorine group Phenyl ring Hydrozone structural feature Thioether group Erythromycin derivative Phosphonate Phosphate

Table 5-6 The combination of tables 5-3 to 5-5 indicates that there are a lot of variables which contribute to 38% of the variation in the dataset. There are a number of variables which appear in tables 5-3 to 5-5 more than once. These variables are carboxylic acid groups, primary and secondary amides and aromatic/enamine groups.

It is clear from table 5-6 that the features which give the most variation to the data set are those listed other variables. These variables are not common in every product. It seems possible to use the information in table 5-6 to help determine how an API may be cleaned from a process plant post manufacturing. This is because there are some differences in the variability of the data set. In order to further investigate the information in data set 1, the Loading plot was analysed .The Loading plot is analysed and discussed in section 5.3.8.

5.3.8 The Loading Plot for Database 1

The Loading Plot of variables for PC1 and PC2 in database 1 (figure 5-9) indicates functional groups and structural features, and their position in relation to each other.

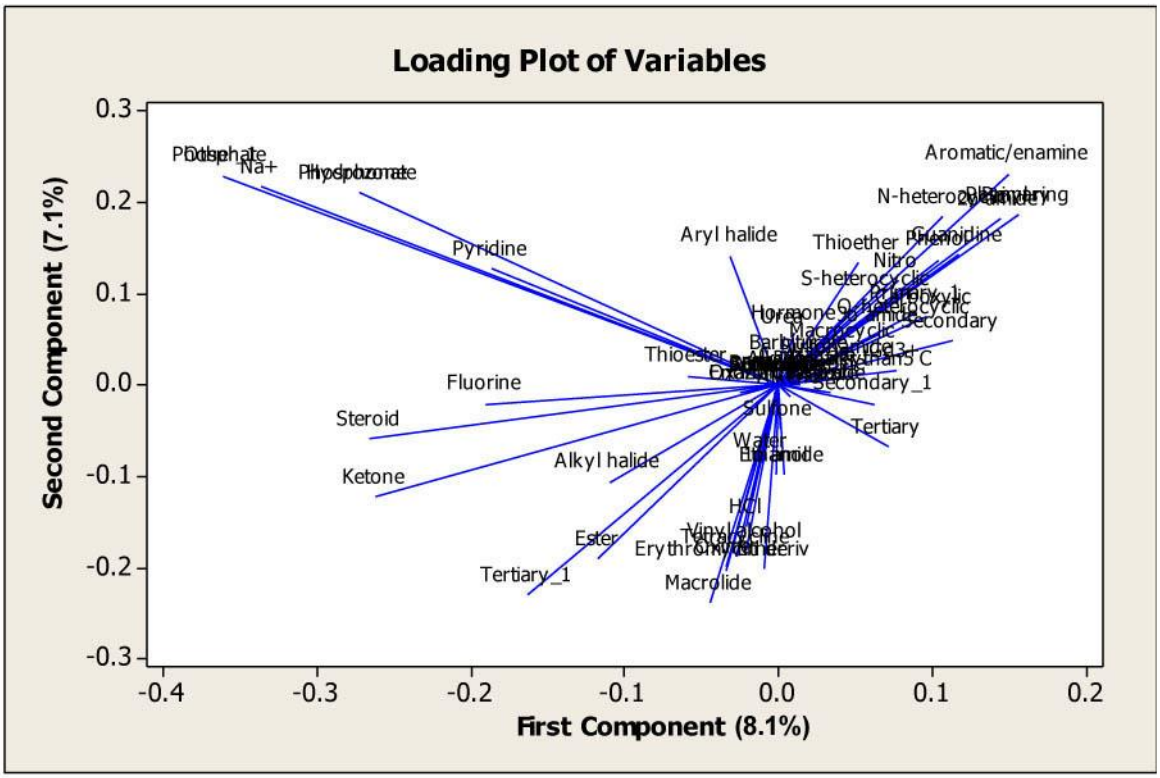


Figure 5-9 showing the Loading plot of all variables for the principal components PC1 versus PC2.

Figure 5-9 was the only Loading plot considered during this analysis and it shows the loadings or variables used in the analysis (the functional groups and structural properties) in relation to the eigenvalues from the first and second principal components. In order to show how this information corresponds with the scree plot (figure 5-3) and the score plots (figures 5-4 to 5-9) it is necessary to show the loading plot with groupings identified with circles (figure 5-10). The groupings were identified by visual inspection of the Loading plot.

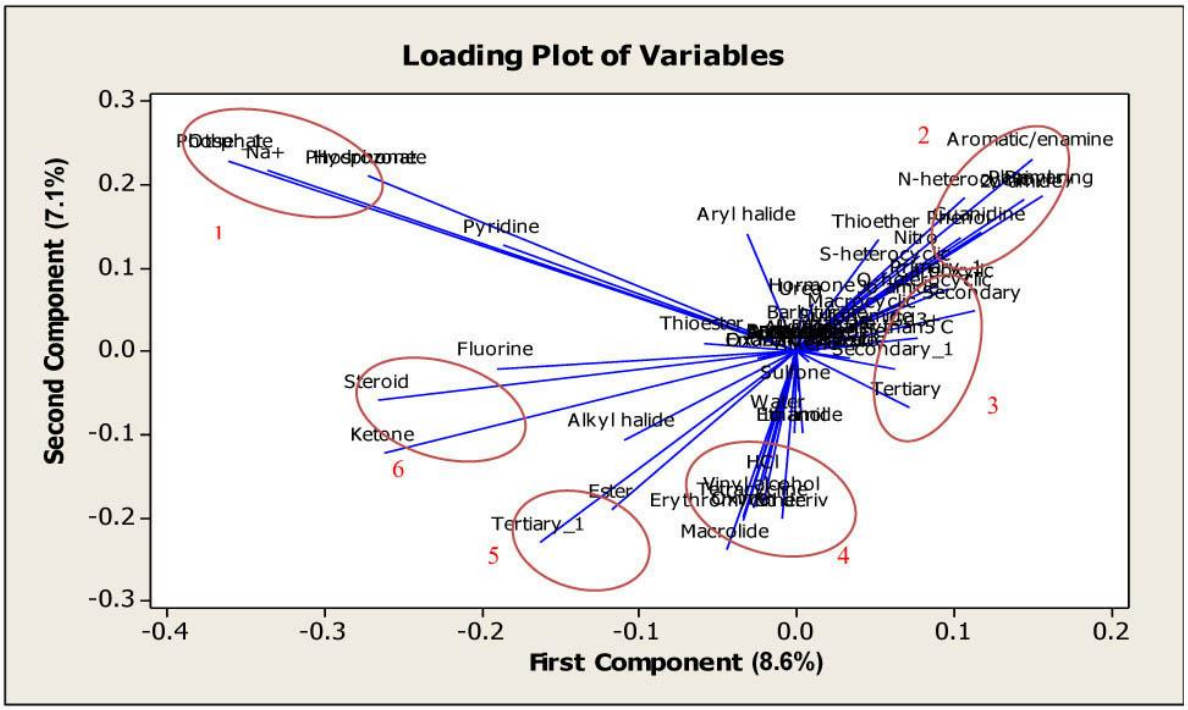


Figure 5-10 Loading plot showing relationship between first and second component. The circles drawn on the figure indicate groupings or points of interest.

Figure 5-10 shows groups of interest which have been circled and numbered one to six. Overall, the loading plot shows some properties which are separate to the main cluster of points which is central to zero on the plot. The circled and numbered variables are the ones of interest for clarity these are given in table VII (appendix V).

The significant characteristics and individual points found in the analysis of figure 5-10 are described as follows -

Group 1

Group 1 contains characteristics of importance as determined by the scree and score plot analysis. This cluster contains phosphate and phosphonate functional groups, Na+ associated groups and Hydrozone characteristics. In addition this group contains a characteristic which refers to forms of acid other than Sulfonated and Carboxylic. In this group it is possible to determine that there are both functional groups which are considered soluble in water (Na+ associated groups) and less soluble in water (phosphate, phosphonate functional groups) and the structural characteristic hydrozone.

Group 2

Group 2 contains primary characteristics Aromatic/enamine, Phenyl ring, Secondary amide. It also contains secondary characteristics, functional groups and structural features N-heterocyclic structures, Phenol (moderately soluble in water) and Guanidine. All of these functional and structural features are water soluble.

Group 3

Group 3 contains primary characteristics Secondary amine, Gd3+ and Tertiary alcohol. It also contains Tertiary amine. Carboxylic acid also associated in this cluster but not as distinctly. Group three lies close to the main cluster of information, which is not considered of importance. All of these functional and structural features are water soluble.

Group 4

Group 4 contains the primary characteristic Macrolide. In addition to this Ether, Erythromycin derivatives, plus other secondary characteristics are identifiable. All of these characteristics are moderate to lowly soluble in water.

Group 5

Group 5 contains the primary characteristic Tertiary alcohol functional groups. It is presented in a group on its own in this interpretation but it is associated with Ester functional groups on the loading plot. Tertiary alcohol groups are soluble in water. If the hydrocarbon chain length of the alcohol increases the functional group becomes less soluble. Esters are soluble in water but if the chain length increases the solubility of the ester decreases.

Group 6

Group 6 contains two characteristics which are considered of secondary importance (therefore coloured blue in table IV in appendix V). These are Steroid organic frameworks and Ketone (Carbonyl) functional groups. These groups appear to be significantly distinct from the rest of the information, due to their position on the loading plot. Both steroid organic frameworks and ketones are soluble in water. These two characteristics lie close to Fluorine on the loading plot.

In addition to the above identified groups several other characteristics appear distinct from the rest of the information. They include Pyridine, Fluorine, Thioester, Ester Aryl halide and Alkyl halide.

Within the dataset on the loading plot it is difficult to determine the position of some characteristics, which are considered of importance in the analysis of the score plots (figures 5-4 to 5-10. These variables are Primary amine groups and Vinyl alcohol functional groups. Conclusions from the analysis of all of the information from database 1 will be discussed in the next section 5.3.9.

5.3.9 Dataset one analysis conclusions

The information provided in sections 5.3.3 to 5.3.8 considers the analysis of the data set in database one which contains information on functional groups and structural features of API's identified as being manufactured by Britest members. The three different plots, the scree plot (figure 5-3), the score plot (figures 5-4 to 5-9) and the loading plot (figure 5-10) all give information about the variables of importance in the first six principal components. Considering all of the information and interpreting it is the best way to ensure that a good proposal or model is generated in order to give Britest members an idea of how information on functional and structural properties of chemical products can be used to devise cleaning strategies.

The information generated from these data has drawn the following conclusions.

There appears to be some characteristics within the dataset which can be considered of higher importance when clustering information together. This information could be separated into primary and secondary characteristics. These were determined on the basis of the number of factors - the loadings plot information (section 5.3.8), the score plot information (section 5.3.7) and the scree plot (section 5.3.4) which determined how many principal components should be considered of importance for this analysis. The primary characteristics are composed of the following structural and functional group characteristics grouped by classification (table 5-7 and table 5-8).

Primary Characteristic					
Amine	**Alcohol OH**	**Acid**	**Carbonyl Groups**	**N Groups**	**Other characteristics**
Primary amine Secondary amine Aromatic/ enamine	Tertiary alcohol structure Vinyl alcohol	Carboxylic acid	Secondary amide	Oxime	Phosphonate Phosphate

Table 5-7 Identified primary characteristics functional groups.

Primary Characteristics	
Organic Framework	**Framework Features**
Hydrozone Phenyl ring Macrolide	Na+ Association Gd3+ Association

Table 5-8 Identified framework and structural primary characteristics.

Primary characteristics have been identified which could be used to link or group chemicals together, in order to suggest similar cleaning methods, which was the aim of this research (table 5-7 and table 5-8). In addition to this, a series of secondary characteristics (table 5-9) were identified that in combination with primary characteristics help could define pharmaceutical products into categories for realising different cleaning methodologies.

Secondary Characteristics			
Alcohol	**Carbonyl**	**Other**	**Organic Framework**
Secondary alcohol	Ketone Ester	Ether	Steroid

Table 5-9 Secondary characteristics of importance.

In addition to the information presented in tables 5-7, 5-8 and 5-9, combinations of the primary and secondary characteristics of interest in grouping products are shown in Table 5-10.

Group Number	Characteristics
Group 1	Na + Association, Hydrozone, Phosphate, Tertiary alcohol structure Secondary alcohol structure, Ketone, Aryl Halide, Steroid
Group 2	Tertiary alcohol structure, Macrolide, +/-Ketone, Ester, Ether Vinyl alcohol
Group 3	Aromatic /enamine, Secondary amide, Phenyl ring, Phenol Alkyl >5 carbons, Guanidine, Primary alcohol
Group 4	Oxime group, +/- other properties
Group 5	Aromatic/ enamine
Group 6	Tertiary amine
Group 7	Secondary amine, Carboxylic acid
Group 8	Phenyl ring, +/- Secondary amine

Table 5-10 Combinations functional groups and structural features of interest as a basis for cleaning methodology development, based on the score plot information.

It is considered that the features identified in tables 5-7, 5-8, 5-9 and 5-10 will help identify potential groupings or links that allow cleaning methodologies to be tailored for particular uses and cleaning challenges. This was one of the aims of this research. The research in this report has served to give an indication of certain pharmaceutical products which could potentially be grouped together for cleaning purposes.

The research carried out in this report means that there is now an identifiable list of chemical functional groups and structural features that can be provided to Britest members. What is not known is whether certain functional groups or structural characteristics are genuinely more important than other characteristics. If many of the identified features are present in a pharmaceutical product it is not possible to determine which ones are dominant.

In order to decide if analysis on data set one is the correct set of data to use as a model to begin to understand how to cleaning manufacturing equipment more effectively by understanding the fundamental aspects behind cleaning, it is important to consider the data in the second data set database two. Database two contains information relating to the same API's considered in database one but uses a different approach. This is to determine the effectiveness of beginning to analyse the data by considering the physicochemical properties as a basis to cluster the API's, and determine if cleaning can be carried out on the basis of this. Section 5.4 begins to look at the analysis of this second data set comprising of information on physicochemical properties.

5.4 Database Two Analysis

5.4.1 Introduction

This section discusses analysis of database two which contained information on the physicochemical properties of the API's chosen for this study (listed in appendix III). As previously mentioned it was difficult to obtain this data due to the nature of the information required. Data of a physicochemical nature is often not published by companies due to confidentiality. In addition, it was often not possible to obtain specific data as it was just not available. Full characterisation of API's to the extent which was required for this research is not carried out. Therefore, in order to analyse this database it was necessary to reduce the number of API's used (to 55) removing those where information was not available.

5.4.2 Database two analysis Scree plot examination

Database two was examined by PCA as discussed in chapter 4. Initially, the scree plot was examined (figure 5-11).

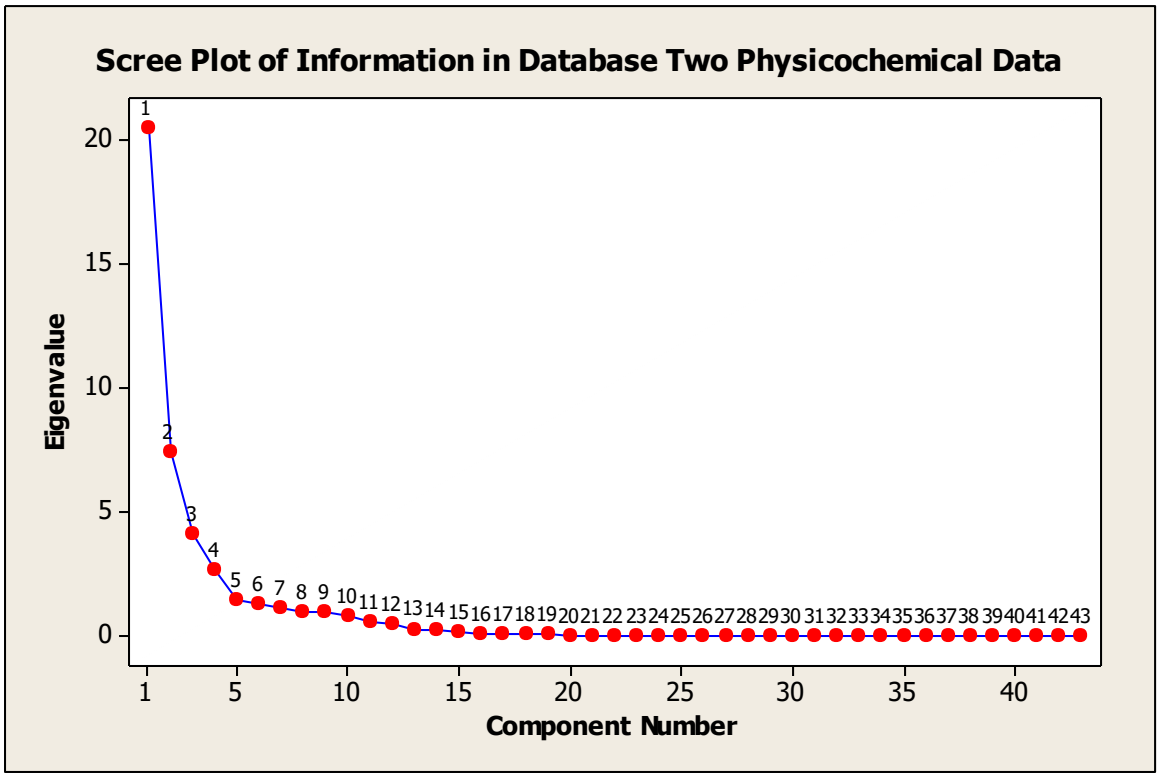


Figure 5-11 Scree plot of physicochemical property information found in database 2.

The scree plot (figure 5-11) gives a visual plot of eigenvalues against principal component numbers. The number of components which contributed to the most variability was determined from this plot. These were deemed the significant principal components. Figure 5-11 shows a typical scree plot shape as described by Minitab (version 16). There were a number of components essential to the variability of the data. It was determined from figure 5-11 that the 'elbow' point of the data was up to five components. Within this dataset there were four components with an eigenvalue of greater than two. These data points significantly contributed to the variation, accounting for 80.4% across the data. It was therefore determined that the first four principal components were the ones to focus on in order to analyse both the score and loading plots. Principal component 1 had the greatest total variation in the data set with an eigenvalue of 20.443 and it accounted for 47.5% variation in the data. The second principal component had an eigenvalue 7.390 and accounted for 17.2% of variation in the data. The third principal component had an eigenvalue 4.095 and accounted for 9.5% of the data variation. The fourth principal component had an eigenvalue of 2.661 and accounted for 8.04% of the data variation.

The remaining components contributed 20% of the variation and were not considered for analysis. This decision was supported by the fact that components five and six only accounted for 6.7% of the variation within the dataset.

In order to establish what this meant in terms of the data involved, it was necessary to examine the variable values for each principal component one to four. This was carried out by examining the data for each given variable and determining its significance. The data was studied and eigenvalues below -0.150 and above 0.150 were determined as cut off values of significance. The reasoning and method used for database 1 (section 5.3.3) was used for database 2. This was carried out in order to give continuity throughout the analysis of all data in this research. It is possible to determine which components contributed to the variation in database 2. The scree plot showed which of the variables contributed to the variation within the data set considering the first four principal components (table VIII, appendix V).

The list of variables, which showed the greatest variability in the data set within the first four principle components is given in Table VIII, appendix V. The variables analysed which appeared to give the most variation across the four principle components were tPSA and Polar surface area. These variables were also linked in the dendrogram analysis described in section 5.2 of this chapter. For reasons discussed in section 5.2 it is not surprising that these variables should add considerably to the variation in this data set. Other variables showed significant contributions to variability across the initial four principle components (table VIII, appendix V). Variables including H bond acceptors, ACD/KOC (pH5.5) and ACD/BCF (pH5.5), contributed to a large amount of the variability within the data set. These variables were considered to cluster in the dendrogram analysis (section 5.2). Other variables adding to the variability included exact mass and molecular weight, Fluorine and Nitrogen content (both gases at room temperature and therefore considered easy to remove from equipment during cleaning). Also contributing to the variability in the data set in the first 4 principal components was Log P, which was identified in section 5.2 in the dendrogram analysis as relevant to cleaning research, as it is an indicator of chemical solubility. Surface tension and vapour pressure showed a high amount of variability within the first four components. These variables were not identified in the dendrogram analysis as adding greatly to the variation in the data set. Surface tension could have been identified because it is known to be important in cleaning processes. Surface tension is the tangential force that keeps a fluid together at an air liquid interface and it is considered an important factor in the choice of cleaning agents (Durkee, 2014).

The only variables which were not accounted for in the first four principle components were Oxygen and Sulphur. It was considered that this was because both of these variables were present in most of the products used in the analysis.

In order to relate this analysis back to the pharmaceutical products it was necessary to establish which products had the properties identified in table VIII (appendix V). The best method considered to determine the API's of interest was to examine the information in the scree plot in association with the score plot. This was carried out in section 5.4.3.

5.4.3 Database two information: Score plot analysis

In addition to the analysis carried out on the scree plot (section 5.4.2), the score plot was analysed for the first two principal components. Analysis of the score plot (figure 5-12) showed the relationship between the scores (chemical pharmaceutical products). This gave an indication of the relationships between the components determined by the first two principle components.

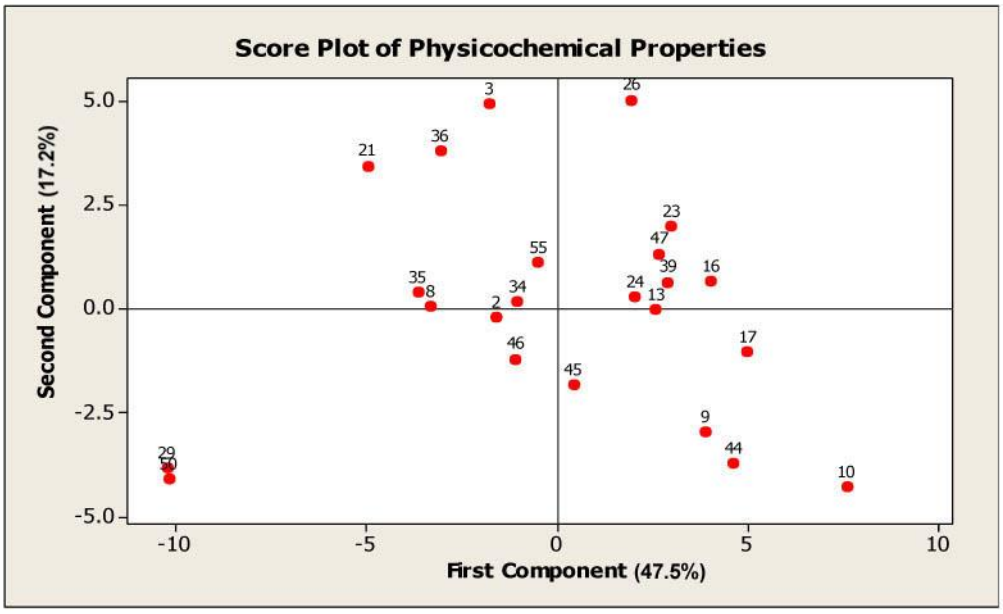


Figure 5-12 Score plot of Physicochemical information in database 2.

In Figure 5-12 pharmaceutical products were shown located around the zero point on both axes. Clusters of components were prominent on the score plot. These were represented by numbers which refer to different pharmaceutical products. In order to visualise the groupings and individual points which require analysis, the score plot was reproduced in Figure 5-13 with indications of groupings shown by circles. These groupings and individual points were identified by visual analysis of the score plot.

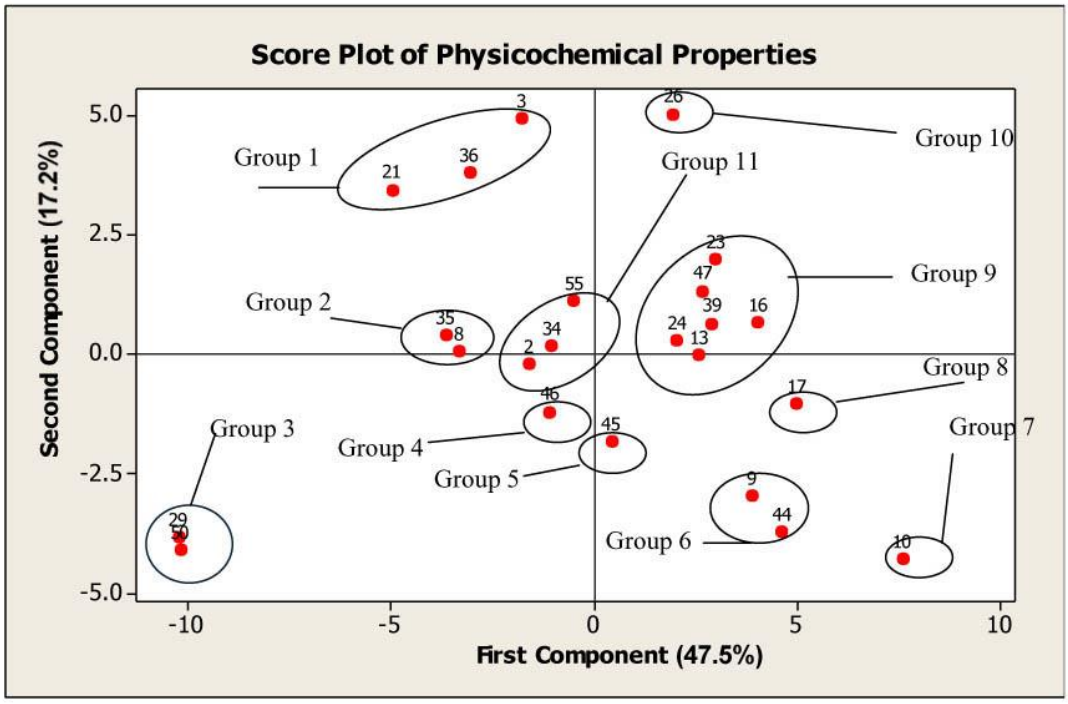


Figure 5-13 Score plot of First and Second Component taken from analysis of database two, physicochemical information. The circled and numbered red dot numbers in each group refer to the pharmaceutical product reference number in the analysis.

As figure 5-13 shows there were several identifiable groupings and prominent features circled. These were identified in table IX in appendix V and are discussed below.

Figure 5-13 gives products that stand alone as points of interest. For ease of interpretation these are defined in this analysis as groups. The content of each group is identified as follows and this includes the number assigned to the product during the analysis (This helps identification on figure 5-13). These were groups 4 (Progesterone (45)), 5 (Plendil (44)), 7 (Ciclesonide (9)), 8 (Fluticasone propionate (17)) and 10 (Hytrin (25)). There were several groups of two chemicals which were group 2 consisting of Meperidine (34) and Brofen (7), group 3 consisting of chemicals Isoflurane (28) and Severane (49) and group 6 consisting of Calcijex (8) and Paricalcitol (43). Two groups contained 3 chemicals these were group 1 containing Atenolol (3), Meprobamate (35) and Gabapentin (20) and group 11 containing Warfarin (54), Marcaine (33) and Androgel (2). There was one group of six chemicals which was group 9. This contained chemicals Gopten (22), Quinapril (46), Halobetasol (23), Mometasone furoate monohydrate (38), Clobetasol propionate (12) and Dexamethasone dipropionate (15). The remaining chemicals within the data set were not found on the score plot. Table IX in appendix V gives chemicals which were identified by the score plot of first two principal components. Observing the data in table IX (appendix V) and in figure 5-13,

one relationship between the grouped products becomes clear. The chemicals have grouped primarily according to their molecular weight among other factors. This observation is clear when shown on a plot (figure of the molecular and exact mass of each product shown in table IX (appendix V). The relationship between the products is subject to other factors which has been determined because several of the products with similar molecular weights have not grouped together. In order to analyse the information given in figure 5-13 further, the common or unique features were determined for each group (table 5-11). This analysis was carried out by comparing non normalised data for each variable.

Group Number	Identifying features
1	In group 1 the chemicals Atenonol, Meprobamate and Gabapentin were identified. The similar physicochemical properties in the group were the fact that the chemicals all have the same elements present which are Carbon, Oxygen, Hydrogen and Nitrogen. All three chemicals had a low C Log P value ranging between -0.66 and 0.915, which was the lowest of all identified groups. This group also had low ACD/LogD (pH5.5), ACD/LogP (pH5.5), ACD/BCF (pH5.5), ACD/Log D (pH7.4) values. Additionally, surface tension values were similar and H bond donor ability had a tendency to be higher in this group than the other identified groups.
2	Group 2 chemicals were identified as Meperidine and Brofen. Both of these API contain no Fluorine, Sulphur or Chlorine. Meperidine had Nitrogen present but Brofen does not. The chemicals both had a similar Henry's Law value, a similar C Log P value and a similar CMR value. They had the same number of freely rotating bonds (4). They had a similar value for Index of Refraction and Surface Tension and a similar Boiling point.
3	This group of API consisted of 2 chemicals which were Isoflurane and Severane. These chemicals both contained no Sulphur or Nitrogen but Isoflurane contained Chlorine. The API in this group contained the lowest Gibbs Energy and Henry's Law values identified among the data set. Similar characteristics in this group between the two chemicals were a low Heat of Form value and the same value for tPSA. They also had similar C Log P values, CMR values, Vapour pressure values, Enthalpy of vaporisation, Density, Polarisation value, no H bond acceptors, and similar ACD/Log D (pH7.4) values. Both API had similar Boiling point, and Surface Tension and Molar Volume values. The Index of Refraction was also similar and the number of Freely rotating bonds was the same (2). The number of H bond

Group Number	Identifying features
	acceptors was the same and the ACD/ Log D value (pH5.5) and ACD/Log P values were similar.
4	Groups 4, 5, 7, 8 and 10 only contain one chemical each. It is therefore not possible to compare the common physicochemical features in these groups.
5	
7	
8	
10	
6	Group 6 contained two API's these were Calcijex and Paricalcitrol. These chemicals had very similar physicochemical characteristics. These included the same exact mass, the same molecular weight and the same number of Carbon, Oxygen and Hydrogen molecules. Both chemicals had no Fluorine, Sulphur, Nitrogen or Chlorine molecules. The chemicals had similar boiling points, melting points, critical temperature values, and critical pressure values. The API have similar Log P numbers, MR values, Henry's Law values, similar Heat of Form values and the same tPSA values. Calcijex and Paricalcitrol had similar CMR values, ACD/Log P (pH5.5) and ACD/Log D values. The ACD/BCF (pH5.5) values were very similar and also higher than those of the other groups, with the exception of group 7. The ACD/KOC (pH5.5) values were higher in this group than in all other groups. Both chemicals had 3 H bond acceptors and a higher number of freely rotating bonds than most other groups identified. The API's had similar molar volumes and boiling points and the same flash point values. High ACD/BCF (pH7.4) and ACD/KOC (pH7.4) values were indicative of this group. The group of chemicals also had the same number of H bond donors, the same polar surface volume, similar molar refractivity values, similar polarizability values and Enthalpy of vaporisation values
9	Group 9 was the largest group of chemicals identified. It contained chemicals Gopten, Quinapril, Halobetasol, Mometasone furoate monohydrate, Clobetasol propionate and Dexamethasone dipropionate. The chemicals in this group were identified by the fact that they have a very similar number of Carbon, Oxygen and Hydrogen atoms. None of the chemicals in this group had Sulphur present and the amount of Fluorine, Nitrogen and Chlorine varied between the chemicals. The API's had a similar boiling point which was higher in this group than in the other groups

Group Number	Identifying features
	(with the exception of group 10). The chemicals in this group also had similar critical pressure values, critical pressure values, and critical volume values, Gibbs Energy values and Log P values. Heat of form values in this group were similar and all of these values were negative. The chemicals had similar tPSA values and CMR values
11	In group 11 there were three chemicals. These were Warfarin, Marcaine and Androgel. Similar physicochemical characteristics between the three chemicals were the amount Carbon present. All of these chemicals had no Fluorine, Sulphur or Chlorine atoms. One chemical, Marcaine had Nitrogen present. The chemicals had similar melting points, MR values, CMR values, ACD/Log P values and molar volumes. The API's are all able to donate one H bond and they had a similar molar refractivity value.

Table 5-11 Features of groups identified in the score plot during PCA analysis of Database 2.

In order to further analyse the data generated in the score plot (figure 5-13) it was compared against the average physicochemical property values for the dataset. A discussion of this analysis is given in appendix V. Appendix V also includes a breakdown of the main physicochemical characteristics of each of the API's identified in table 5-11, and a flow chart indicating the simplest way to determine which group each API would be associated with.

In order to further analyse the information provided during the PCA it was important to examine the score plot of the third and fourth principal components (figure 5-14). This shall be carried out next.

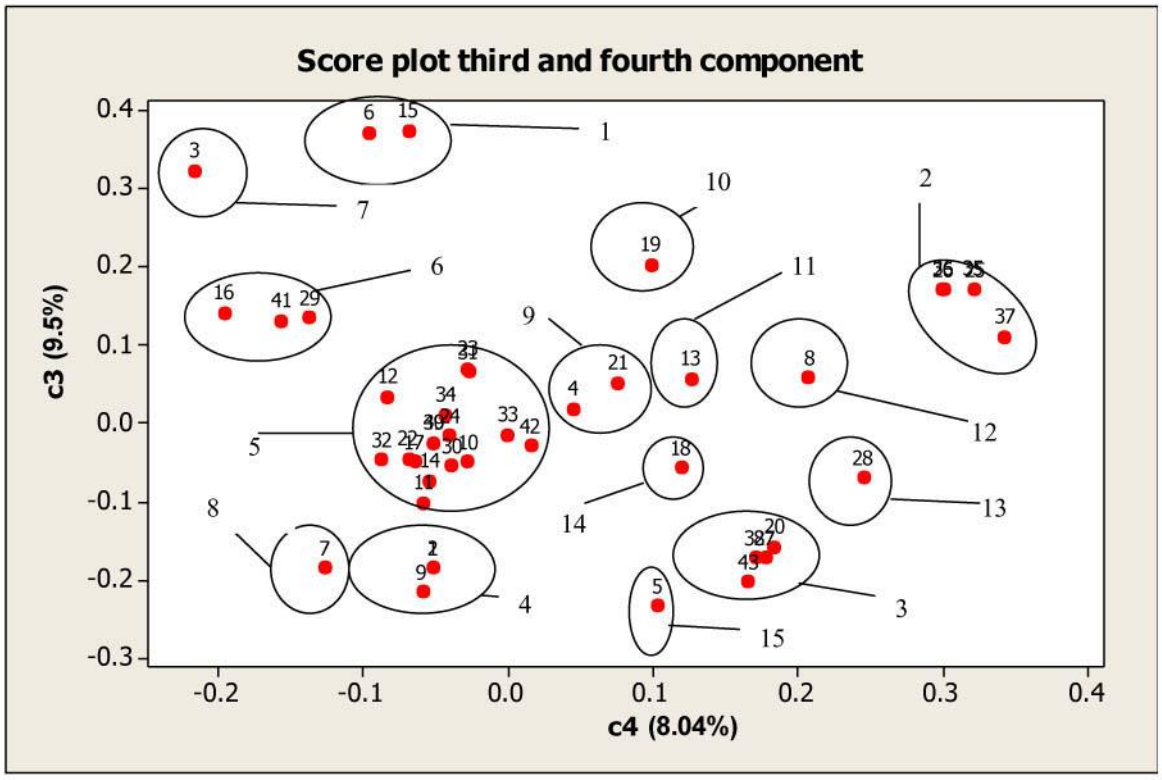


Figure 5-14 Score plot of the third and fourth principal components. Where c3 and c4 indicate principal components 3 and 4.

Figure 5-14 indicated visibly identifiable clustering of variables. The groupings were shown as numbered clusters (1-15) on figure 5-14. An explanation of what each cluster represents was discussed as follows. The numbers given in brackets following a variable indicate the analysis reference number.

Group 1

This cluster showed the relationship between identified variables H (6) and Gibbs energy (15) which were both identified in the third principal component as contributing to the variability.

Group 2

This cluster shows variables ACD/KOC (pH5.5), (26), ACD/KOC (7.4), (36), ACD/BCF (pH5.5), (25), ACD/BCF (pH7.4), (35) and H bond donors (37). These variables are identified in the scree plot analysis as adding to the variation in the first four principal components.

Group 3

This cluster contains variables including tPSA (20), H bond acceptors (27), Polar surface area (38), Vapour pressure (43), F (5). The variables in this group have not been previously identified as contributing to the variation in the score plot of principle components 1 and 2.

They were identified with the analysis of the loading plot, which will be discussed later in this research thesis (section 5.4.4).

Group 4

Group 4 was a small cluster of three variables which were exact mass (1), molecular weight (2) and Cl (9). These variables had been identified as adding significantly to the variation in the scree plot and the loading plot.

Group 5

Group 5 was the largest cluster of variables identified on figure 5-14. It contained variables ACD/Log P (23) not identified as being of significant on any other plot analysed. Carbon (3) identified in both the scree plot and the score plot of principal components 1 and 2. Critical temperature [K] (12) identified in the scree plot as being significant in principal component 1. ACD/Log D (pH7.4), (34) was identified in the scree plot as being significant in terms of adding to the variability within the data set and also in the loading plot, as well as the score plot of the first two principle components. ACD/Log D (pH5.5), (24) had been identified in the scree plot and the loading plot as being of significance. The variable boiling point °C (33) was identified in this plot as being of significance; it was also identified in the scree plot. Variable enthalpy of vaporisation was identified in the scree plot as adding significantly to the variation (42), variable Boiling point [K], (10), was identified as adding significantly to the variation of the data set in the loading plot and the scree plot. Molar volume (30) was found to add to the variability of the data set in both the scree plot and the loading plot. The variable melting point [K], (11) was found to add to the variation of the data set significantly in the scree plot. The variable flash point was found to add to the variation of the data set in the scree plot (32). The variable CMR was found to be significant in the scree plot (22). Variable Molar refractivity (39) was found to be significant in the scree plot. The variable MR (17) was found to be significant in the scree plot.

Group 6

Group 6 contains a cluster of three variables. These were LogP (16), and Index of Refraction (29), both identified in the scree and loading plot as adding to the variability of the data set. Density (41) was identified in the scree plot as contributing to the variability of the data set.

Group 7 to group 15

Group 7 contained one variable which was the variable Carbon (3).

Group 8 contained one variable of interest which was the variable Sulphur (7).

Group 9 contained two variables which were the variable Oxygen (4) and the variable ClogP (21).

Group 10 contained one variable of interest which was heat of form (19)

Group 11 contained one variable of interest which was critical pressure (13).

Group 12 contained one variable of interest which was Nitrogen (8).

Group 13 contained one variable of interest which was freely rotating bonds (28).

Group 14 contained one variable of interest which was Henry's Law (18).

Group 15 contained one variable of interest which was Fluorine (5).

A summary of how this information contributed to the variation in the data set was discussed after consideration was given to the information provided in the loading plot which was interpreted in section 5.4.4.

5.4.4 Database Two: Loading plot analysis

The loading plot showed the relationship between functional groups and structural properties of the data set in relation to the eigenvalues for the first and second principal components. In order to determine how this information corresponded with the scree and score plot, it was necessary to show the loading plot with groupings identified. For clarity, variables of interest were indicated by circles labelled A to L (figure 5-15). These were considered as follows -

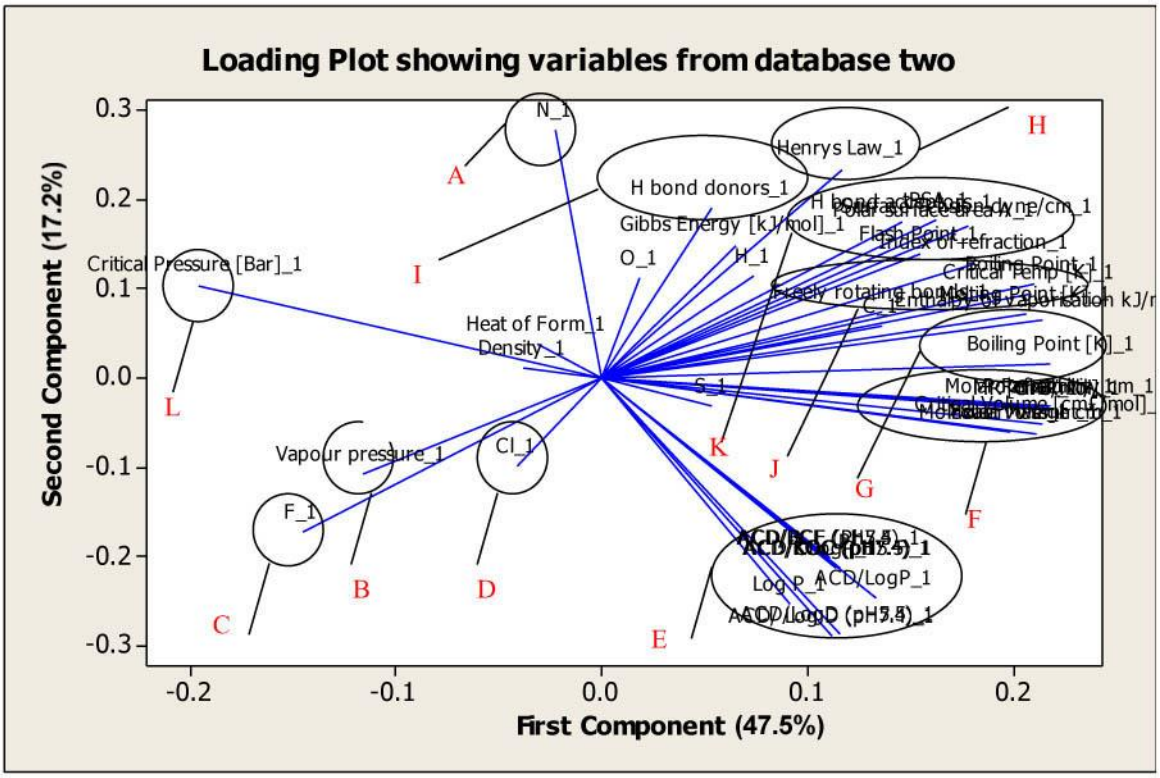


Figure 5-15 Loading plot showing the relationship between the first and second component. The circled data indicates groups of interest. These were labelled A to L.

Variables of interest are shown circled and labelled A – L on figure 5-15. These were discussed as follows -

Group A

This group contained one physicochemical feature which was the element Nitrogen. This element was found to be of significance in the scree plot and the score plot.

Group B

Contained one feature of importance which was the physicochemical characteristic of vapour pressure identified in the scree plot.

Group C

Fluorine was the physicochemical feature identified of significance on the loading plot in group C. This was also found to be significant in the scree and the score plot.

Group D

The element Chlorine was found to be of significance in this group on the loading plot. Chlorine was also found to be significant in the scree and the score plot.

Group E

This grouping of physicochemical characteristics on the loading plot contained all data relating to ACD/ BCF, ACD/KOC and ACD/ Log D at both pH values (pH 5.5 and pH7.4). There is a relationship between these values (as previously discussed in section 5.2) and therefore they would be expected to be close on the loading plot. Group E contained features which were shown to be of significance on both the scree and the score plot. The other physicochemical characteristic present in this grouping was the Log P value which was found to be significant in the scree plot analysis.

Group F

This grouping contained features which were identified as being of significance in data analysis of the scree and the score plot. In this group both molecular weight and exact weight were present. Molar volume identified in the scree plot analysis was also present in this group.

Group G

This group contained the boiling point [k] but not boiling point °C. There is a very definite relationship between these two values and it is not known why the analysis showed them to be different. Boiling point [k] was found to be significant in the score plot.

Group H

This was a large group of variables which contained features that are considered significant by scree plot analysis, such as polar surface area and H bond acceptors. There were other physicochemical characteristics in this group which were not identified as having significance in the scree or the score plot. This included the variables of flash point and index of refraction.

Group I

This group contained H bond donors which have been identified as of significance during analysis of the score plot.

Group J

Group J contained variables which were identified in other plots. These variables included critical temperature, melting point, freely rotating bonds, boiling point (°C), enthalpy of vaporisation. This group also included Carbon.

Group K

Group K contained the variables flash point, index of refraction, surface tension dyne, H bond acceptors, Polar surface area and tPSA.

Group L

This group contained only one variable which was Critical Pressure (Bar). This was not identified as being significant by any other plot.

Other Observations

The characteristics density and heat of form are shown as distinctly separate (location wise) from other features on the loading plot.

PCA of database two indicated that there were a number of interesting clusters and linkages in the data. This information was collated in table 5-12.

Variables of Significance in Database Two determined by Principal Component Analysis			
Molecular weight	Critical Temperature	C	ACD/LogD (pH7.4)
Exact Mass	Critical Pressure	F	ACD/BCF (pH7.4)
Gibbs Energy	Critical Volume	H	ACD/KOC (pH7.4)
MR	H bond acceptors	N	ACD/LogP
Henry's Law	Freely rotating bonds	Cl	ACD/LogD (pH5.5)
Heat of Form	H bond donors		ACD/BCF (pH5.5)
Flash Point	Index of Refraction		ACD/KOC (pH5.5)
Boiling Point (°c)	Molar volume		Log P
Boiling Point [K]	Surface Tension		tPSA
Melting Point	Polar surface area		CLogP
	Molar refractivity		CMR
	Polarizability		
	Density		
	Enthalpy of vaporisation		
	Vapour pressure		

Table 5-12 Variables contributing to the greatest variability in database two.

The variables identified in the analysis of database two are shown in table 5-12. A majority of the information show in table 5-12 relates to factors which are important when considering cleaning vessels post manufacturing API's. This includes a number of the elements found in the data set (C, F, H, N, Cl). This could be due to the nature of the elements which can determine the state of the API at room temperature gas, liquid or solid at given temperatures. Fluorine, Hydrogen, Nitrogen and Chlorine are gases at room temperature. The number of

Carbon (C) atoms in an API related to the size of the product and potentially the solubility. Henry's Law relates to the solubility of a gas in a liquid and therefore if the state of a product is considered an important characteristic. It is not surprising that this variable has also been identified. Molecular weight, relative mass (MR) and exact mass have been identified in this analysis as being of significant within the data. These relate to the size of a molecule which is a factor in cleaning. Variables which relate to the states at different temperatures were also identified in this analysis as being of greater significance than other variables. These include flash point, melting point and boiling point. These physicochemical characteristics are considered important when designing cleaning protocols. Gibbs free energy was found to be of significance in this research. This may be because it concerns the amount of free energy associated with chemical reactions that can do work. Heat of Form is of significance as it is the amount of heat generated during the formation of one mole of a compound from its component elements. The importance of H bonds in relation to cleaning has been discussed in section 5.2. Surface tension, tPSA and Polar surface area and their importance in cleaning were also discussed in section 5.4.2. Variables including ACD/LogD (pH7.4), ACD/BCF (pH7.4), ACD/KOC (pH7.4), ACD/LogP, ACD/LogD (pH5.5), ACD/BCF (pH5.5), ACD/KOC (pH5.5), Log P, CLogP and their significance to cleaning has already been considered in this chapter.

Both databases of information were examined and the analysis has been discussed. It is important to now consider what PCA analysis of the full data set indicates. This requires analysis of databases 1 and 2 together (both the functional and structural detail of the API's and the physicochemical information collected on the same group of API's). This will be discussed in section 5.5.

5.5 Database 3 Analysis

Both databases analysed by PCA have indicated patterns and connections in the data which could be used to indicate appropriate cleaning methodologies in the pharmaceutical industry. It is thought that analysis of the two databases together as one database may indicate more groupings and patterns than the individual databases alone. Analysis of database three by PCA will be discussed in this section. Initially, the analysis will focus on the information contained in the Scree Plot (section 5.5.1).

5.5.1 The Scree Plot

The scree plot (figure 5-16) visualises the eigenvalues for each principal component during analysis of Database 3.

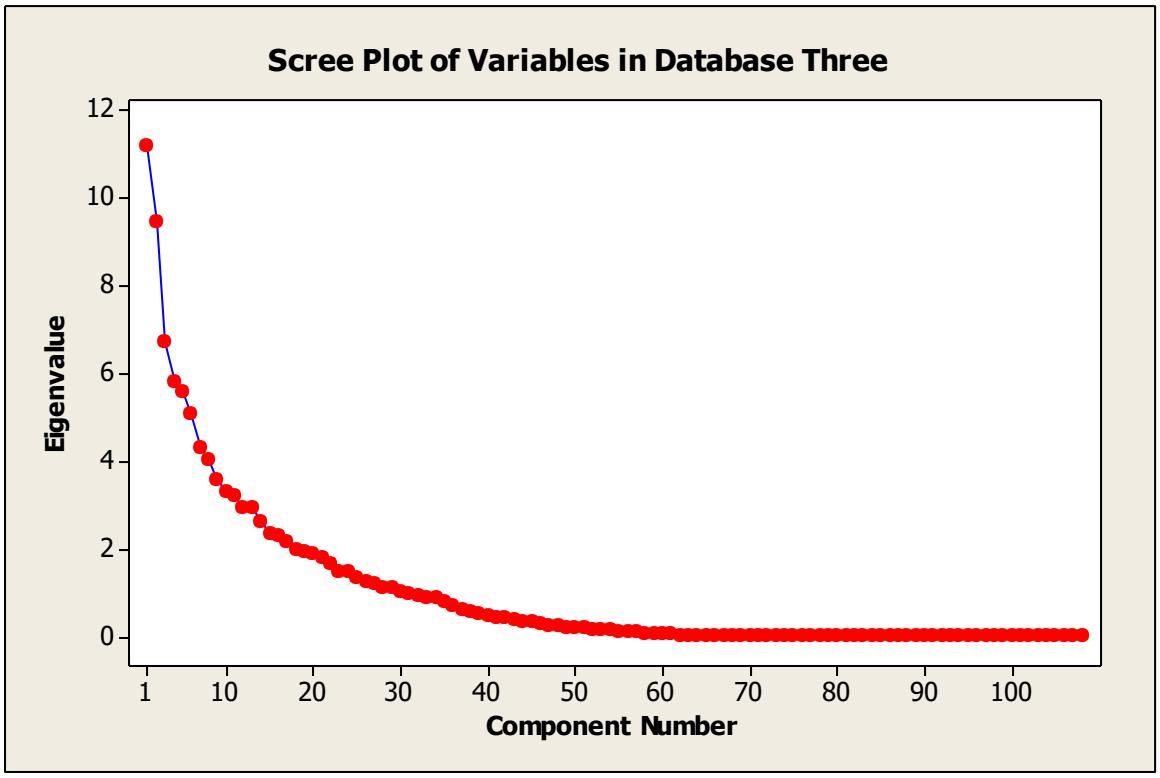


Figure 5-16 Scree Plot indicating the eigenvalues given for each component during analysis of Database 3.

Figure 5-16 showed the relationship between the eigenvalues and the components. The scree plot shows the typical shape, as already described earlier (section 5.3). It was important to determine the number of principal components which added the most to the variability in the data set. The first 14 components within the data set gave the most variability in the data. This number of components corresponded with the “elbow” point on the plot. The scree plot also shows that there is a considerable amount of data which contributes to less than 1% of the variability in the data set. This includes data from the last 49 principal components. Principal components 1 to 14 account for 65.3% variability in the data and therefore the first 14 principal components were examined. The eigenvalues from the first 14 principal components were found for each variable (table XII, appendix V). Some variables were not represented in the first 14 principal components. These variables, including functional and structural features and physicochemical properties, were listed (table XIII, appendix V) and not considered further in the analysis of the information presented by the scree plot for database three.

The number of variables represented in the first 14 principal components was large (Tables XII and XIII). It was not considered appropriate to consider them all to be of significance at this stage of the analysis. It was important to reduce the number of variables at this stage of the analysis to include only those contributing the most to the variability in the data set. This

was carried out by reducing the number of principal components considered to be of significance to the first six components. A further observation from the analysis of this data set was that the majority of the variables of interest in the analysis were physicochemical characteristics of the APIs. The information provided in table XIII also indicates that most of the variables not of interest in the first 14 principal components were functional and structural features.

Table 5-13 was constructed in order to reduce the number of variables considered to significantly add to the variability of the data set. This indicates the variables which add the most variability within the first six principal components.

Variable name Functional and structural features	**Principal component associated with the variable**	**Variable name Physicochemical features**	**Principal component associated with the variable**
Aromatic/enamine	c2	Nasal and inhalation classification	c4 c5
Primary 1	c6	Injectable classification	c4 c5
Tertiary 1	c5 c6	Antibiotic classification	c4 c5
Ketone	c3	API classification	c4 c5
2 amide	c2 c6	Exact mass	c2 c6
Tertiary amide	c2	Molecular weight	c6
Ether	c6	Contains N	c2 c3
Thioether	c6	Contains P	c4
Fluorine	c5	Contains Na	c4
Pyridine	c4	Contains I	c6
Aryl halide	c4	Boiling Point [K]	c1
Alkenes	c5	Melting Point [K]	c1 c3
Phosphonate	c4	Critical Temperature [K]	c1 c3
Hydrozone	c4	Critical Pressure [Bar]	c3
Other features	c4	Critical Volume	c1

Variable name Functional and structural features	**Principal component associated with the variable**	**Variable name Physicochemical features**	**Principal component associated with the variable**
		(cm3/mol)	
Phosphate	c4	Gibbs Energy (KJ/mol)	c5
Nitro	c6	MR (cm3/mol)	c1
Steroid	c3 c5	Henrys Law	c1 c3
S-heterocyclic	c6	tPSA	c3
		C Log P	c1
		CMR	c1 c3
		ACD/Log P	c1
		ACD/Log D (ph5.5)	c1 c3
		ACD/BCF (pH5.5)	c1 c5
		ACD/KOC (pH5.5)	c1 c5
		H bond acceptors	c2 c3
		Freely rotating bonds	c2
		Index of Refraction	c1 c6
		Molar Volume (cm)	c2
		Surface Tension dyne/cm	c6
		Flash Point	c2
		Boiling Point (°c)	c2
		ACD/BCF (pH7.4)	c1 c5
		ACD/KOC (pH7.4)	c1 c5
		H bond donors	c2
		Polar surface area A	c2
		Molar Refractivity	c2

Variable name Functional and structural features	Principal component associated with the variable	Variable name Physicochemical features	Principal component associated with the variable
		(cm)	
		Enthalpy of vaporisation kJ/mo	c2

Table 5-13 Variables associated with the first 6 principal components during analysis of the scree plot for database three.

Table 5-13 showed the number of variables considered important to adding to the variability of the data set within the first 6 principal components. This represented 40.5% of the variability in the data set. The variables of importance are functional group and structural features as well as physicochemical properties. It became clear from table 5-13 that there were more physicochemical variables represented in the first six principal components than in the functional groups and structural features category. It was also apparent that most of the variables which were represented in the first three principal components (accounting for 25.3% of the variation), were physicochemical variables.

In order to understand the analysis of database three it is important to consider the score plot and the loadings plot. Therefore, before considering what the clustering or groupings of data mean for industrialists considering pharmaceutical plant cleaning it was important to analyse the score plot in section 5.4.2.

5.5.2 The Score Plot

The score plot (figure 5-17) was produced during analysis of database three containing all data (of functional and structural features and physicochemical properties of products). During this section it was necessary to consider what the analysis indicated. A number of clusters and points of interest were visually identified on figure 5-17. In order to investigate the points of interest it was necessary to annotate the score plot to identify groupings (figure 5-17).

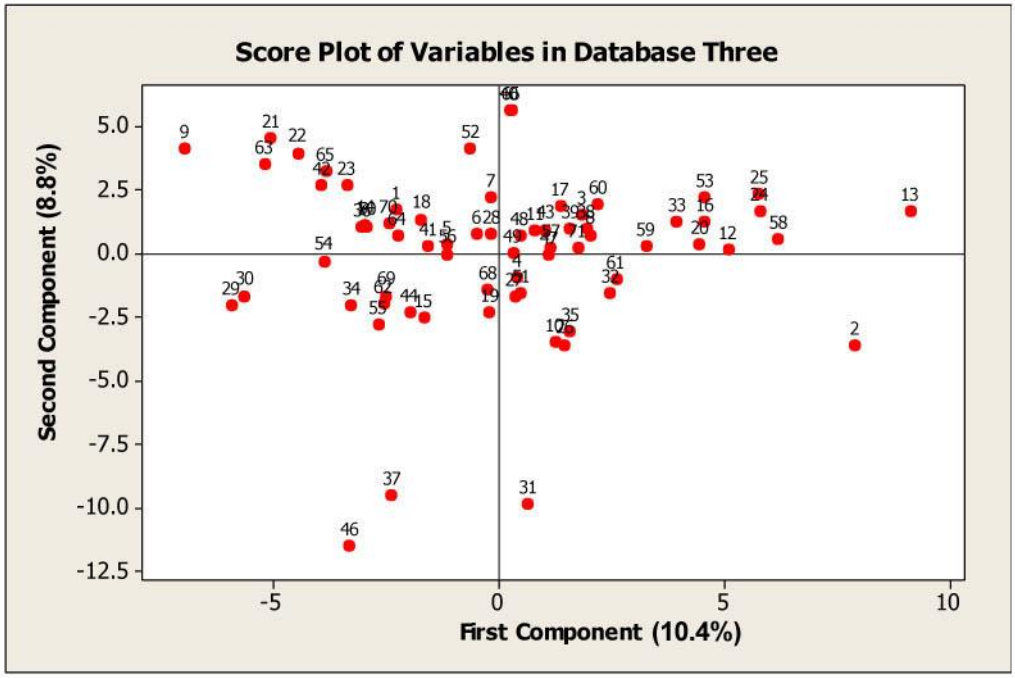


Figure 5-17 Score Plot indicating the relationships between variables in database three. The red dots on the plot indicate a specific API. The identity of the API can be determined by the number it is associated with.

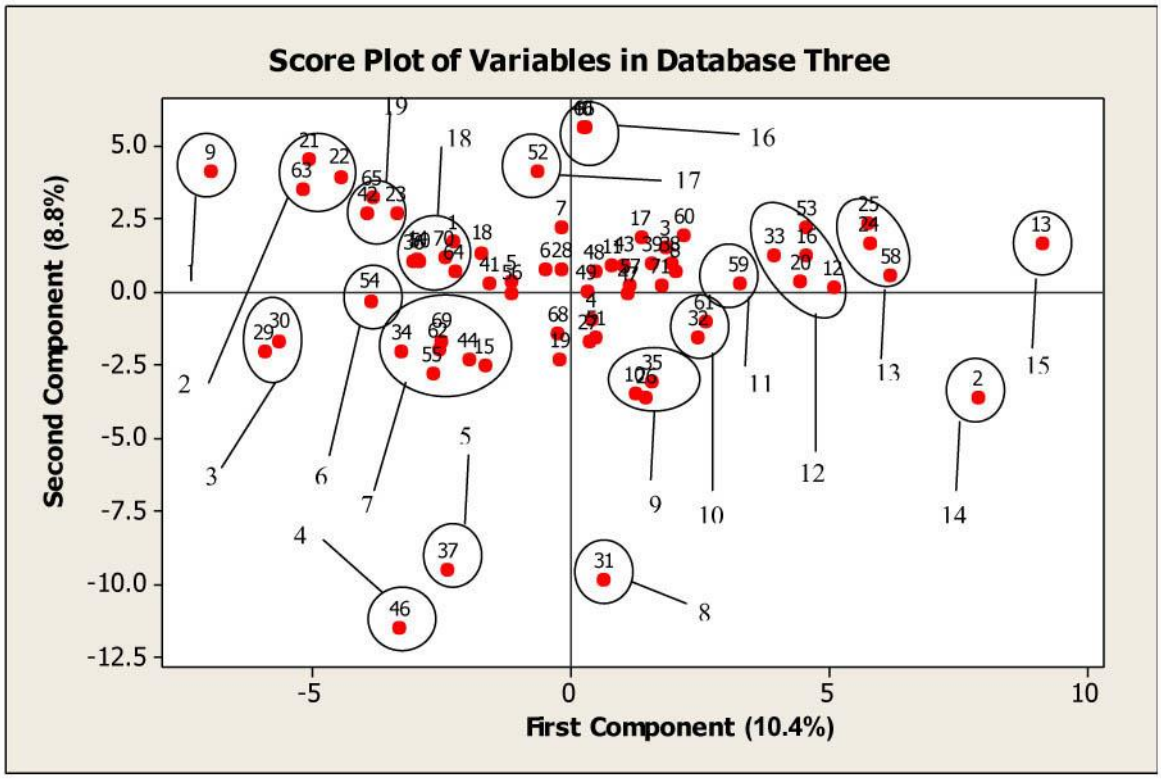


Figure 5-18 Annotated score plot indicating clusters of interest during analysis of Database 3.

In Figure 5-18 points and clusters of interest are indicated. To help analyse these points further the clusters are presented in table 5-14.

Cluster number (figure 5-15)	**Row numbers associated with the cluster**	**Products associated with the cluster or point of interest**
1	9	Betamethasone disodium phosphate
2	21 22 63	Doxycycline hyclate, Doxycycline monohydrate, Roxithromycin
3	29 30	Gadopentetate dimeglumine, Gadopentetate monomeglumine
4	46	Lupron (Leuproreline)
5	37	Iodixanol
6	54	Nimbex (Cisatracurium besilate)
7	34 55 62 69 44 15	HPMPC (Cidofovir), Nizatidine, Ranitidine, Eprosartan (Teveten), Klacid (Clarithromycin), Clarithromycin
8	31	Ciclosporin
9	35 26 10	Hytrin, Folic Acid, Blopress (Candesartan cilextil)
10	61 32	Quinapril, Gopten (Trandonapril)
11	59	Plendil (Felodipine)
12	12 20 16 33 53	Calcijex (Calcitriol), Dexamethosone dipropionate, Clobetasol propionate, Halobetasol, Mometasone furoate monohydrate
13	58 24 25	Paricalcitol (Zemplar), Fluticasone furaroate, Fluticasone propionate
14	2	Aluvia (Lopinavir or Ritonavir)
15	13	Ciclesonide
16	66 40	Severane, Isoflurane
17	52	Mometasone furoate anhydrous
18	64 70 1 50 14 36	Salmeterol xinafoate, Venlafaxine, Advicor (Niacin or Lovastatin), Methohehexital, Citanest (Prilocaine), Imdur (Isosorbide mononitrate)

Cluster number (figure 5-15)	Row numbers associated with the cluster	Products associated with the cluster or point of interest
19	23 65 42	Epival (Sodium valproate), Sevelamer, Ivermectin
Main group	60 18 41 5 56 19 68 4 51 27 6 28 7 47 57 71 49 48 11 43 39 3 17	Progesterone, Cycloserine, Isradipine, Bambec (Bambuterol), Olanzapine, Deflox (Terezosin hydrochloride), Tamsulosin, Atenolol, Metolazone, Furosemide, Beclomethasone dipropionate, Gabapentin, Beclomethasone dipropionate monohydrate, Marcaine (Bupivacaine), Oxis (Formoterol), Warfarin, Meprobamate, Meperidine, Brofen (Ibuprofen), Ketoprofen, Iopamidol, Androgel (Testosterone), Conholip.

Table 5-14 Groupings and points of importance as shown in figure 5-18. API names were shown in the table (alternative names for the same API are shown in brackets proceeding the initial name of the API).

Table 5-14 shows that there were several similarities between the groups identified during examination of the score plot for Database 1. These included individual products of interest. Betamethasone disodium phosphate (9) and Lupron (46) were found to be physical located apart from other API's on the score plot. Doxcycline hyclate and Roxithromycin were also physically located close to each other on the plot. Other groups identified in table 5-14 were very different from those identified in the score plot of Database 1 examining principal components 1 and 2. This is probably due to the increase in variables. There are not many similarities between groupings on the score plot of Database 2 and Database 3. Ciclesonide is located physically apart from the rest of the data on both score plots.

At this point in the analysis it is important to consider the information which can be gained by examining the score plot of principal components 3 and 4 (figure 5-19).

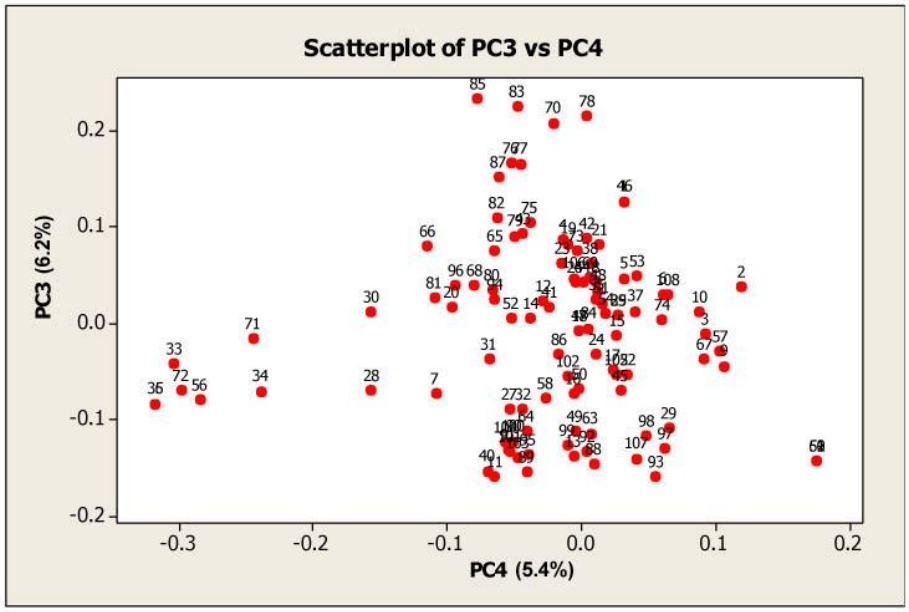


Figure 5-19 Scatter plot of principal components 3 and 4 taken from principal component analysis of database three.

Several points of interest among the variables can be found in Figure 5-19. These were examined further using an annotated diagram to determine groups and points of interest (Figure 5-20).

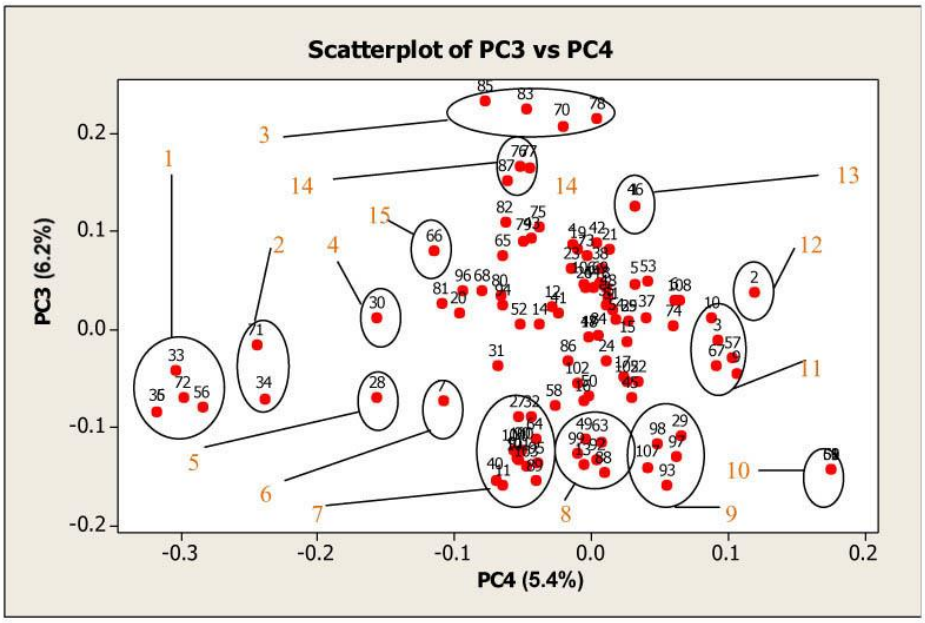


Figure 5-20 Annotated figure 5-19 showing points and clusters of interests. The red dots indicate a variable of interest which was identified by the number associated with it in black writing. Clusters were indicated by the circled data and given a group number in orange writing.

Table 5-15 shows clustering of variables according to the relationship between principal components 3 and 4. Figure 5-20 indicates there were a number of points of interest which correspond to variables. These variables are listed in table 5-15 according to the groups on figure 5-20.

Groupings identified in Figure 5-20	**Variables associated by row number in analysis**	**Variables associated in the identified group by name**
1	33 72 35 36 56	Phosphonate, Na association, Other features and groups, Phosphate groups, Na atoms
2	71 34	Phosphate atoms Hydrozone
3	85 83 70 78	tSPA, Henrys Law, Nitrogen atoms, Critical Pressure (bar)
4	30	Aryl halide
5	28	Pyridine
6	7	Tertiary amide-1
7	11 40 89 103 95 91 101 64 27 32 104 90	Ketone groups, Alkyl groups greater than 5 carbons, Steroid features, Molecular weight, Fluorine groups, ACD/Log D (pH5.5), ACD/BCF (pH5.5), ACD/KOC (pH5.5), Molar Volume cm, ACD/KOC (pH7.4), Polar Surface Area A, Molar Refractivity cm.
8	49 63 99 13 92 88	Macrocyclic, Exact Mass, ACD/Log D (pH7.4), H bond acceptors, ACD/LogP, Ester groups
9	29 98 97 107 93	Alkyl halide groups,

Groupings identified in Figure 5-20	Variables associated by row number in analysis	Variables associated in the identified group by name
		Boiling Point, Flash point, Enthalpy of vaporisation kJ/mo, Freely rotating bonds.
10	59 60 52 61 62	Nasal and inhalation, Injectables, Antibiotics and API classifications, Barbiturates
11	10 3 67 57 9	Carboxylic groups, Tertiary, Phenol groups, Gd3+ groups, Fluorine atoms
12	2	Secondary amide
13	46	Phenyl ring structures
14	77 76 87	Melting Point [K], Critical Temperature [K], CMR
15	66	Oxygen atoms

Table 5-15 Variables associated with principal components 3 and 4 generated during analysis of database 3.

Table 5-15 indicates that there are a number of variables which were considered important in the analysis of the principal components 3 and 4. Prior to further discussion of these findings it is important to determine what principal components 5 and 6 can show in terms of analysis. Figure 5-21 was created using the principal components 5 and 6.

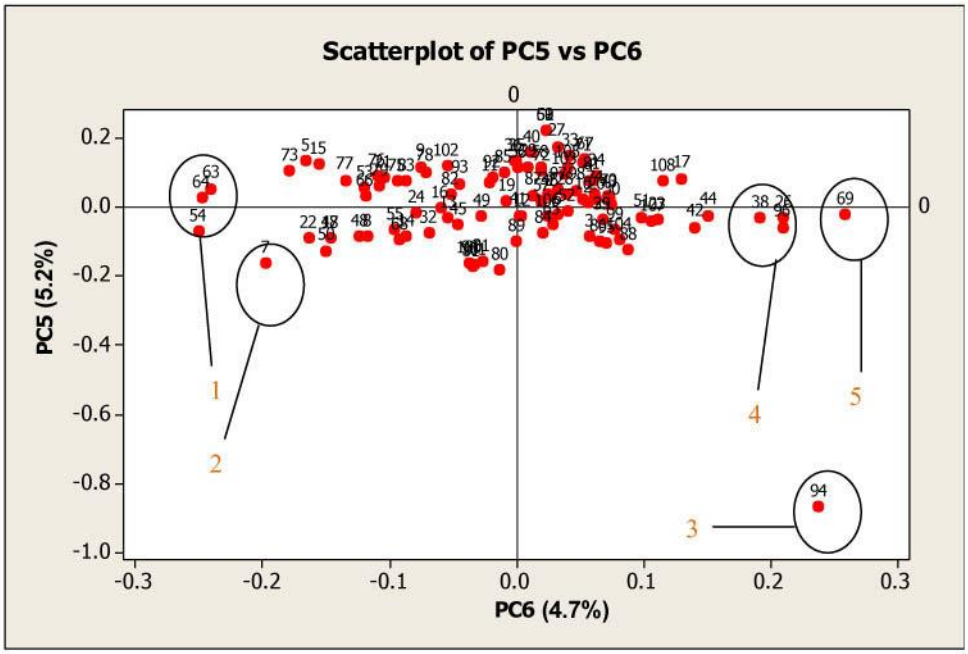


Figure 5-21 Annotation of figure 5-20 indicating the groups and points of interest. The red dots indicate a variable of interest which was identified by the number associated with it in black writing. Clusters were indicated by the circled data and given a group number in orange writing.

Figure 5-21 shows that a majority of the variables are located around the zero axes. It was possible to identify points of interest which were marked with numbered circles on the plot. The variables associated with the groupings were shown in table 5-16. The most interesting point on figure 5-21 was the variable 94 which was data related to index of refraction. This was because it was located at some distance from the other variables.

Groupings identified in Figure 5-20	Variables associated by row number in analysis	Variables associated in the identified group by name
1	54 64 63	Ethanol, Molecular weight, Exact mass
2	7	Tertiary amide
3	94	Index of Refraction
4	38 26 96	Nitro groups, thioether groups, Surface Tension dyne/cm
5	69	Sulphur molecules

Table 5-16 variables of interest in groups identified from the scatterplot of principal components 5 and 6 while analysing Database 3.

Analysis of figure 5-21 indicated that there were several variables which were located distinctly away from the main data set. Table 5-16 lists these variables. Several variables are located close enough to visually group together. These are group 1 and group 4. Group 1 contains variables ethanol, molecular weight and exact mass. As exact mass and molecular weight are strongly linked these two variables would be considered likely to group together. It is not known why ethanol should also cluster in this group. In group 4, the variables nitro groups and thioether groups were clustered together with surface tension. Nitro groups are groups known to draw electrons away from a reaction centre. Thioethers are volatile functional groups. The connection between the two functional groups and ethanol is not clear. For ease of explanation individual points of interest on figure 5-21 are described as groups in this research, therefore individual points of interest are labelled as groups 2, 3 and 5. Group 2 contains the characteristic tertiary amide, which is known to show low solubility in water. It is not known why the physicochemical characteristic index of refraction is physically distinct from other variables. Sulphur atoms are the only variable in group 5 which are physically located close to group 4 on figure 5-21. This could be because there is a link between Sulphur atoms and Nitro groups (sulphur groups are found in nitro groups). It is possible to state that there was not a lot of information gained in examining 5-21. The next stage of the analysis will focus on examining the loading plot for Database 3 in section 5.5.3

5.5.3 Loading Plot Analysis

To complete the analysis of Database 3 it is important to consider the information which can be gained by examining the Loading Plot (figure 5-22). Figure 5-22 shows the relationships between the variables when the first and second principal components are plot.

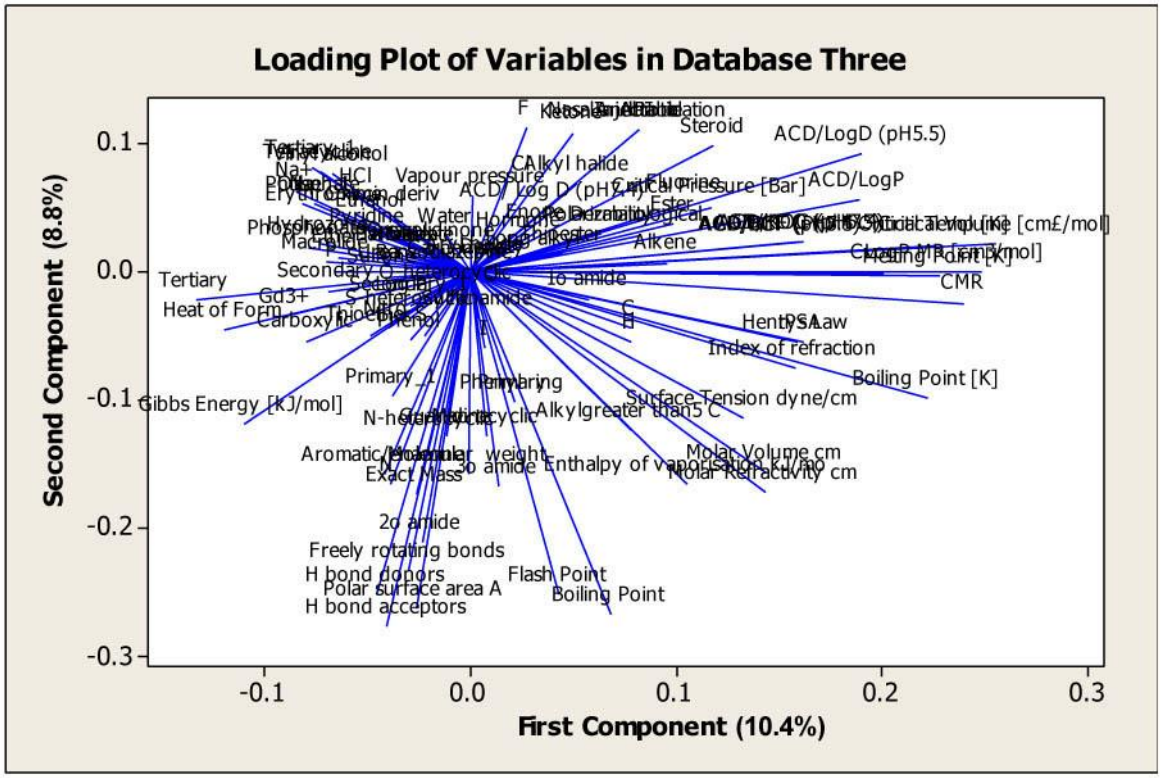


Figure 5-22 Loading Plot indicating the relationship between the variables in database three.

Figure 5-22 shows the variables and indicated by the distance from the zero axes the importance of some variables to the variation of the data. For clarity the clusters of significance are shown on an annotated figure of 5-23. The variables were clustered according to visual inspection and the clusters thought to be of significance are shown in table 5-17.

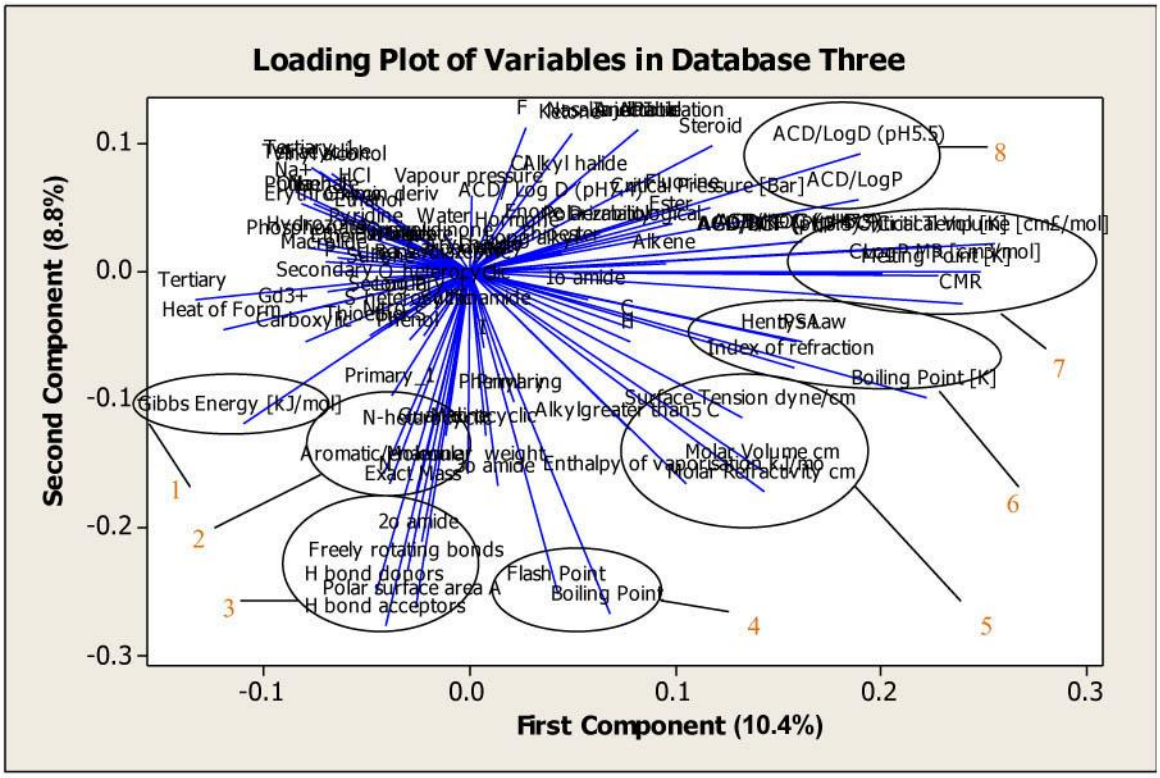


Figure 5-23 Annotated figure 5-22 showing points and clusters of interests. Clusters were indicated by the circled data and given a group number in orange writing.

Visual examination determined 8 clusters. These are listed in table 5-17.

Groupings identified in Figure 5-23	**Variables associated in the identified group by name**
1	Gibbs Energy [kJ/mol]
2	Aromatic/enamine groups, Exact mass, Molecular weight, N-heterocyclic groups, Tertiary amide, Nitrogen molecules
3	Secondary amide groups, Freely rotating bonds, H bond donors, Polar surface area A, H bond acceptors.
4	Flash point, Boiling Point (°c)
5	Surface Tension dyne/cm, Alkyl greater than 5 carbon, Molar volume, Enthalpy of vaporisation kJ/mo, Molar Refractivity cm.
6	Henry's Law, tPSA, Index of Refraction, Boiling point [K]

7	CMR, Melting Point [K], Mr [cm3/mol], Clog P, Critical Volume, Critical Temperature, ACD/BCF (pH5.5), ACD/KOC (pH5.5), ACD/KOC (pH7.4), ACD/Log D (pH7.4),
8	ACD/Log D (pH5.5), ACD/LogP

Table 5-17 showing variables identified on the loading plot (figure 5-22).

Table 5-17 shows that a lot of the variables determined as contributing to the variability in the individual database analysis also contribute to the variability of the whole data set. Analysis of database three by PCA has indicated a number of variables adding to the variability of the data in each plot (the Scree plot, the score plot and the loading plot). The collated variables of significance in database three are shown in table 5-18 for clarity.

Variables of interest structural features and functional groups		**Variables of interest physicochemical properties**	
Aromatic/enamine		Nasal and inhalation classification	CMR
Primary 1		Injectable classification	ACD/Log P
Tertiary 1		Antibiotic classification	ACD/Log D (ph5.5)
Ketone		API classification	ACD/BCF (pH5.5)
2 amide		Exact mass	ACD/KOC (pH5.5)
Tertiary amide		Molecular weight	H bond acceptors
Ether		Contains N	Freely rotating bonds
Thioether		Contains P	Index of Refraction
Fluorine		Contains Na	Molar Volume (cm)
Pyridine		Contains I	Surface Tension dyne/cm
		Boiling Point [K]	Flash Point
Aryl halide		Melting Point [K]	Boiling Point (°c)

Alkenes		Critical Temperature [K]	ACD/BCF (pH7.4)
Phosphonate		Critical Pressure [Bar]	ACD/KOC (pH7.4)
Hydrozone		Critical Volume (cm3/mol)	H bond donors
Other features		Gibbs Energy (KJ/mol)	Polar surface area A
Phosphate		MR (cm3/mol)	Molar Refractivity (cm)
Nitro		Henrys Law	Enthalpy of vaporisation kJ/mo
Steroid		tPSA	
S-heterocyclic		C Log P	

Table 5-18 variables identified as significant in database three.

Examination of table 5-18 and the scree plot of Database 1 showed the following similarities. Variables that were considered to contribute to the greatest variation in both Databases 1 and 3 are listed as follows - Aromatic/enamine, primary amine and tertiary amine, secondary amide, phosphate and phosphonate groups. Structural features included Hydrozone structures. The reason why these variables could be considered to add to the variability in the data set has been discussed in section 5.3. In addition to these variables, other variables which added to the variability in the data set in Database 3 were ketones, ether groups, Thioether groups, Pyridine, Aryl halide groups, Alkenes, a group with other features and Nitro groups. Structural features adding considerably to the variability included Steroid and S-heterocyclic structures.

Every variable identified in Database 3 was common to Database 2 in terms of adding significantly to the variability of the data set with the exception of the product categories. There were some additional variables in Database 2 which were not identified in analysis of Database 3. These were Polarizability, Density, and Heat of Form. The analysis of the three databases has determined a number of variables which add to the variability of the data set in each case. There was a lot of information generated during this analysis. It was therefore

necessary to determine the most useful data to use in order to improve pharmaceutical plant cleaning. This will be considered in the next section 5.6.

5.6 Model creation

In order to examine and make sense of the theoretical results, it was vital to determine how this research relates to industry. Analysis of the databases of information indicated that some of the variables clustered together and some individual variables were distinctly different and added significantly to the variation in the data set according to the scree, score and loading plots. The provision of cleaning agent information by company D enabled the first links to cleaning in industry and information on the variables in the score plots (of principal components 1 and 2) for each PCA in particular. Data on specific pharmaceutical products and their solubility are shown in table IX in appendix V.

The information on cleaning agents was plot onto the score plots (of principal components 1 and 2) from each database analysed. The results are given in figures 5-24 - Database 1 information on chemical functional groups and structural features, figure 5-25 - Database 2 information on physicochemical characteristics, and Database 3 - both sets of data combined. This produced some interesting observations.

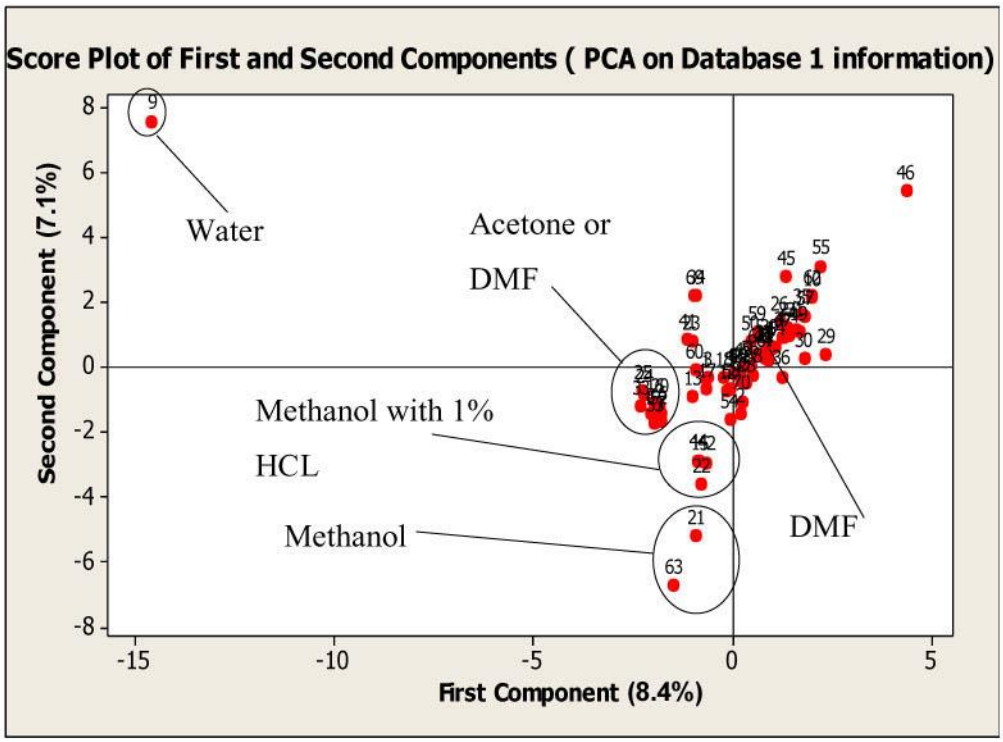


Figure 5-24 Score plot generated during PCA analysis of database 1. The plot has been annotated to show the location of products and the relevant cleaning agent provided by company D. The numbers on the plot by the red dots refer to specific pharmaceutical products.

Figure 5-24 indicates that there is an association between some of the known products and known cleaning agents provided by company D for information provided in Database 1. Table 5-19 shows the variables identified in the pharmaceutical product cleaned from process equipment and the cleaning agent used to clean process equipment post processing.

Identified cleaning agents/ method	Chemical functional groups in identified cluster
Water	Na+ Association, Hydrozone, Phosphate, Phosphonate, Tertiary alcohol association, Secondary alcohol, Ketone, Aryl halide, Steroid
Methanol	Tertiary alcohol structure, Vinyl alcohol, Oxime group, Macrolide, Tertiary amine, Tertiary alcohol, Ketone, Primary amide, Tetracycline, Secondary alcohol, Ester, Oxime, Ether, Erythromycin derivative
Methanol 1% HCL	Macrolide, Tertiary alcohol structure, Vinyl alcohol ,Tertiary amine, Secondary alcohol, Ketone, Ester, Ether, Primary amide, Tetracycline
DMF	Contain a mix of functional groups and identifying features Phenyl Ring, Primary amine, Secondary amine, Tertiary alcohol, Carboxylic acid, Aromatic enamine , Secondary alcohol, Secondary amide, Secondary amide, Primary amide, Ether, Carbamate, N-heterocyclic, Alkene, Alkyl >5 carbons, , Ketone, Oxazolidonone, Tertiary amide, Guanidine, Water, O-heterocyclic, Aryl halide, Sulfonamide, Macrocyclic, Primary alcohol, Tertiary amine, Carbamate, Urea, Barbitute, Thioester, Phenol, Long alkyl, Thioether, Nitro, O-heterocyclic, Sulfonamide, Vinyl alcohol, Phenyl ring, Ketone.
Acetone or DMF	No significant functional group identified by scree plot analysis in any of company D's pharmaceutical products. Do contain some common features Secondary alcohol, Ketone, Ester, Steroid, Alkyl halide, Water, Tertiary alcohol, Ketone, Fluorine, Thioester, Ether

Table 5-19 Variables associated with products produced by company D and the cleaning agents used to remove them from process equipment post manufacture. The black writing indicates features found in the products. Blue writing indicates common features identified in the analysis.

Table 5-19 provides some interesting observations and can therefore be used to state the following in the absence of other practical cleaning information. There was only one product (Betamethasone disodium phosphate) which was freely soluble in water. The product was located distantly from other products on the score plot. There was one product (Doxycycline monohydrate) which was identified as unique (the only product cleaned from equipment using methanol and 1% HCL) in the data provided by company D. Two products manufactured by company D which clustered together are soluble in methanol (Roxithromycin) or freely soluble in methanol (Doxycycline hyclate). The remaining pharmaceutical products were

predominantly in group 11 (Betamethasone acetate, Halobetasol, Dexamethasone dipropionate, Clobetasol propionate, Beclomethasone dipropionate, Beclomethasone dipropionate monohydrate, Fluticasone propionate, Mometasone furoate anhydrous, Mometasone furoate monohydrate) all of these products were soluble in DMF and/or soluble in acetone, which is used in cleaning the products from process plant. There were three company D pharmaceutical products used in the PCA analysis which have not been mentioned so far. Tamsulosin was identified in the main data set. This product is cleaned from plant with DMF according to company D. Products not identified in the analysis on the score plot were Sumatriptan Base which was cleaned from process plants using DMF and Iohexol cleaned from process plants using water. These products were not identified in the analysis on the score plot for database 1. It is not known why Tamulosin was identified and yet a product with the same cleaning agent was not represented.

The above observations may indicate that clustering of products on the score plot. These products have similar functional groups as shown in table 5-19 which may indicate why the same cleaning agents are used to remove them from process equipment. It is important to consider whether this relationship was found on the score plot for PCA analysis of Database 2 (the physicochemical information). This was considered in figure 5-25.

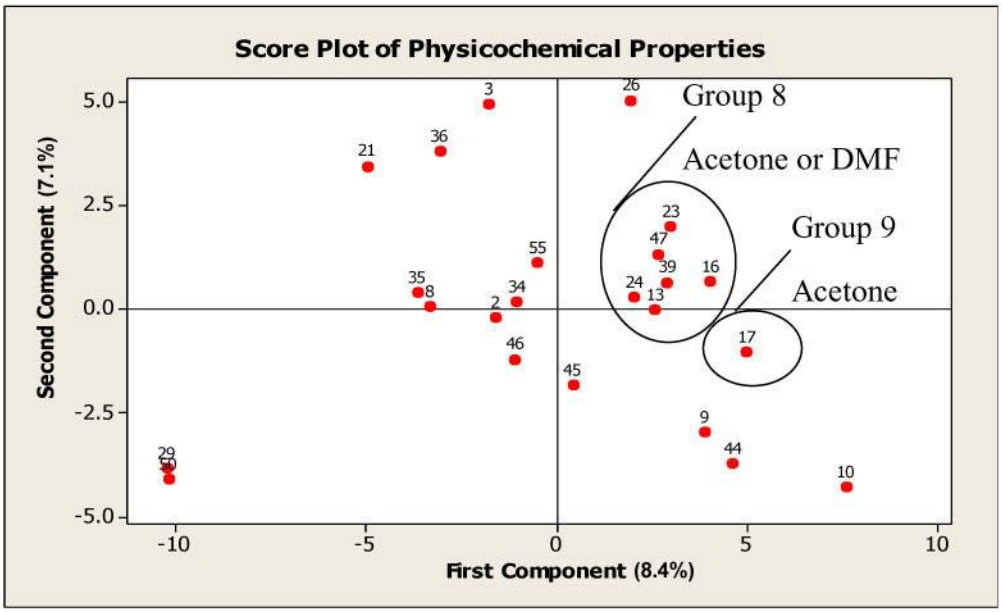


Figure 5-25 Score plot generated during PCA analysis of database 2. The plot has been annotated to show the location of products and the relevant cleaning agent provided by company D. The numbers on the plot by the red dots refer to specific pharmaceutical products.

Figure 5-25 identified only four products made by company D. These products were found in two groups. The first group identified was group 8. This contained one product which was Fluticasone propionate. This product is cleaned from process equipment using acetate.

The second group identified on figure 5-25 contained three products manufactured by company D. These were Mometasone furoate monohydrate, Dexamethasone dipropionate and Clobetasol propionate. Company D uses Acetone or DMF to clean these products from manufacturing vessels. These API's had a similar boiling point, similar critical pressure values, critical volume values, Gibbs Energy values and Log P values. Heat of form values in this group were similar and all of these values were negative. The chemicals had similar tPSA values and CMR values.

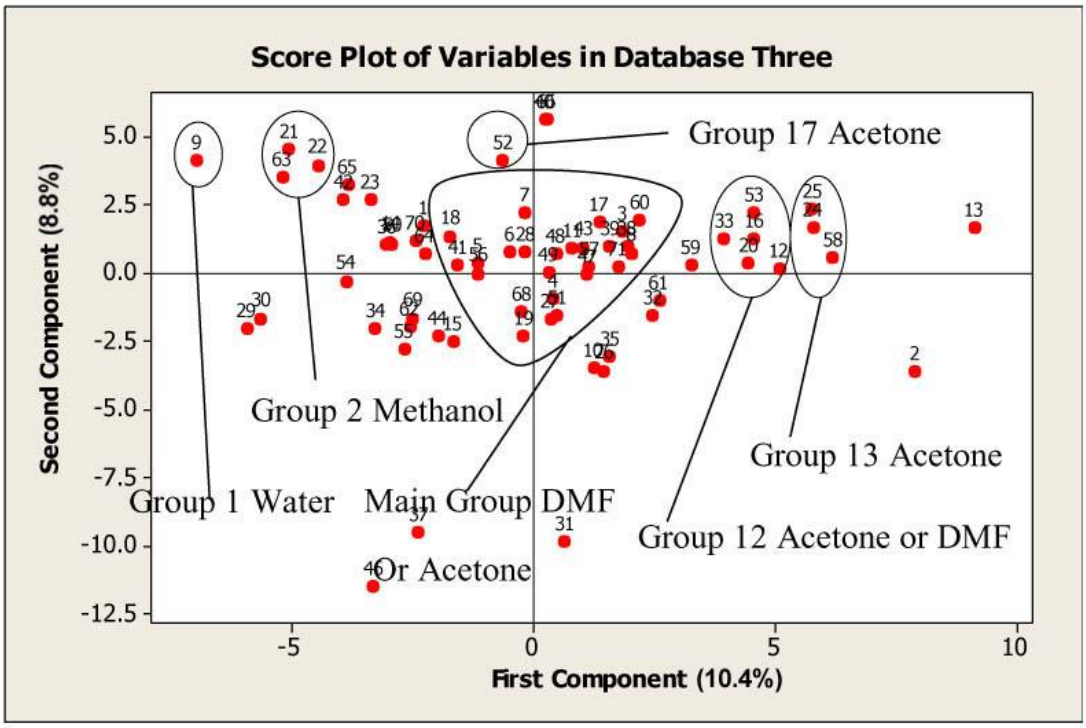


Figure 5-26 Score plot generated during PCA analysis of database 3. The plot has been annotated to show the location of groups of containing company D products and the relevant cleaning agent provided by company D. The numbers on the plot by the red dots refer to specific pharmaceutical products.

Figure 5-26 shows the location of groups which company D products are found and the cleaning agents used to clean the products from equipment post manufacture.

This plot shows Betamethasone disodium phosphate in a distinct position away from other products on the plot. Water was used to clean this product from equipment. Group 2 (found

on figure 5-26) identifies group two as containing company D products Doxycycline hyclate, Doxycycline monohydrate and Roxithromycin. All of these products were cleaned from vessels using methanol, although cleaning of vessels used for manufacturing Doxycycline monohydrate uses a mix including methanol and 1% hydrochloric acid. Group 12 identified on figure 5-26 contained company D products Clobetasol propionate, Halobetasol, Mometasone furoate monohydrate and Dexamethasone dipropionate. All of these products were cleaned from vessels by company D using Acetone or DMF. Group 13 contained company D product Fluticasone propionate, among other products manufactured by other companies. Fluticasone propionate was cleaned from vessels using Acetate. Group 17 contained one company D product which was Mometasone furoate anhydrous. Company D cleaned this product from vessels post manufacturing using Acetone. The final group containing products manufactured by company D was the main data group. This contained three products manufactured by company D. These were Tamsulosin (cleaned from vessels using DMF), Beclomethasone dipropionate and Beclomethasone dipropionate monohydrate (both cleaned from vessels post manufacture by Acetone or DMF).

It was considered important to determine the common features in each of the groups above (chemical functional and structural features and physicochemical features). This was carried out to see if the effectiveness of a cleaning agent could be linked to the variables of each product. This information has already been determined for functional and structural features (Table 5-19). Determining this information for the physicochemical variables was challenging. This was because a lot of the information on variables required to make sense of the analysis was not available. It was decided that using the physicochemical information to interpret why certain chemicals were cleaned from vessels post manufacturing was not viable. The construction of a model to determine the relationship between cleaning agents and functional and structural features is the most reliable and significant use of the analysis. Therefore the model which was used in the rest of the analysis, and used to construct a tool which Britest members can utialise to determine the most effective method of cleaning products from vessel post manufacturing, is the model constructed using database 1 (figure 5-24). The construction of the model has been carried out in this section, but it was important to consider how this model would be used in association with other Britest tools.This will be considered in Chapter 6 which will discuss industrial case studies. The next section (5.7) will provide a summary of this chapter.

5.7 Chapter Summary

This chapter has presented and discussed the results from analysis of the databases of information. This has included the intial dendrogram results on database 2 physicochemical information for each product, which indicated potential groupings of variables and showed that it was possible to determine patterns and clusters in the data. It was considered that using dendrograms to determine patterns in the data was not sufficient to give results necessary to create a tool to help determine the best cleaning method for Britest members. Therefore, another method was used to determine relationships in the data, namely PCA. PCA was carried out on each database of information relating to pharmaceutical products. This suggested that it was possible to find the variables in each data set which added the most variation in each data set. It was possible to determine a number of patterns and clustering effects for each data base during PCA analysis. This revealed a number of variables in each database which were considered of significance. These are listed in Table 5-10 (functional and structural features), Table 5-12 (physicochemical variables) and Table 5-19 (combined functional and structural features and physicochemical variables). The analysis provided information on variables which added to the variability in the each data set and this was related to the pharmaceutical products used in the analysis. During the analysis some of the pharmaceutical products clustered together on plots due to common variables. The link between the information given in the plots and cleaning in industry was key to understanding the data. Information relating to cleaning agents, which was provided by company D, showed a link between certain pharmaceutical products, and their composition, to specific cleaning agents. Using cleaning data from company D, it was possible to determine the best model (data from database 1, 2 or 3) to use to indicate potential cleaning agents. It was found that the best model to use was the score plot (principal component 1 and principal component 2). This was because there was insufficient physicochemical information to give a robust enough model. Therefore the model which was used in case studies was model 1, using the chemical functional and structural data. This model will be applied to industrial data in case studies in chapter 6.

In previous chapters information and results from analysis have been presented with the aim of answering the research questions poised at the beginning of this research in chapter one. In particular research question **RQ2:** What is meant by the term ‘fundamental science’ in relation to process plant cleaning? This chapter has shown it is possible to take information relating to specific products or API’s and analyse it. This identified patterns in the data and indicated the significance of some variables over others in determining the variability in the

data set. Fundamentally, it was possible to say that this information was useful in providing the methodology to answer research question 1, **RQ1:** What would be the best way to increase the fundamental understanding of the science behind cleaning linked to solvent and cleaning agent use.

PCA analysis of databases of information relating to the fundamental composition and structure of pharmaceutical products has shown links to the cleaning agent used for some known products. Without understanding the chemistry behind the variables it is considered that this would not have been possible. In order to establish if this model is a tool which can be useful in industry it is important to consider three points. The first point - is can the model be used in industry to determine potential cleaning methods and help to select a cleaning agent? The number of known pharmaceutical products and successful cleaning agents have been provided by one company. In order to improve the model, it needs populating with more data.

The second point to consider is - do companies use different cleaning agents to successfully clean the same pharmaceutical product from vessels post manufacturing? Knowing this would help Britest members choose the best methodology to clean pharmaceutical plant equipment according to limitations on plant equipment, such as age of equipment, or material of composition.

The third point relates to understanding whole process design. This research has only used information on pharmaceutical products. During manufacturing there will be hundreds of reactions which produce intermediate products and side products. Some of these products will be difficult to remove from process vessels and it is therefore necessary to identify them by using Britest tools such as TM (discussed in Chapter 3). Identification and characterisation of these products can be carried out and PCA could be used to identify them as a cleaning challenge and determine potential cleaning agents to remove them from vessels.

In order to address some of these points the model based on the fundamental science was used with industrial cleaning data provided by Britest members. This was carried in chapter 6.

Chapter 6. Case Studies

6.1 Introduction

The previous chapters have discussed the need for a fundamental understanding of the science behind plant cleaning and how this research could achieve the main research aim RQ1: **What would be the best way to increase the fundamental understanding of the science behind cleaning linked to solvent and cleaning agent use?**

Chapter 5 presented results of analysis, which aimed to answer this research question and improve understanding of industrial plant cleaning. It was considered that this could be achieved by understanding the chemical functional groups and structural features, which compose the pharmaceutical products. The identification and classification of the functional groups in the products used in this analysis resulted in the creation of a model (Chapter 5). It is believed that this model could be used to help predict the best cleaning agent to remove unwanted residues of products from process vessels post manufacturing. In order to demonstrate the importance of the model it must undergo a trial with further industrial data. This chapter examines two case studies to determine if the model developed in this research is able to help determine the best cleaning agent to use to remove a specific product from equipment post manufacturing. In addition to examining cases studies consideration will be given to other Britest tools, which may be useful in understanding other aspects of cleaning as discussed in Chapter 3. The next section discusses the use of Britest tools to solve cleaning challenges, by considering a suite of tools called FUSE (Fundamental Understanding of Science and Engineering).

6.2 FUSE

FUSE aims to consider all aspects of industrial plant cleaning, bringing WPU to cleaning. Chapter 3 introduced the Britest tools and methodologies, which could be used to understand, identify and solve cleaning challenges. In this section it is important to reconsider them with reference to the model, which was created in section 5.6. One of the most important aspects of the research project was to consider cleaning as a part of the manufacturing process. Information provided by Britest members in a survey in chapter 3 indicated that this was not currently the case. The importance of plant cleaning during a manufacturing process is often overlooked and under considered. This research project aimed to give Britest members a way to change how they think about cleaning by realising that it needs to be considered at the beginning of a manufacturing process. In this respect the model developed should be used

before the product is manufactured as a part of Whole Process Design (chapter 1). Therefore, it is thought that the best use of the model (and the data surrounding the model determining functional groups common to products in groups cleaned by certain cleaning agents) is at the beginning of product and process design. Currently the scope or operational space of the Britest tools used to give WPU lies within the blue boxes (figure 6-1).

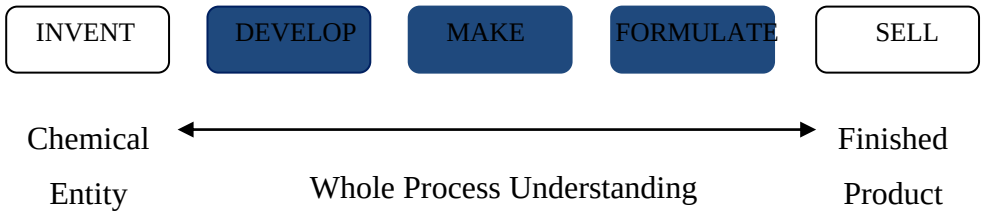


Figure 6-1 Britest tools and methodologies operation space in industrial processes (adapted from Britest material, 2011).

The tool developed during this research should initially be used during the development stage, once the chemical entity is invented. This represents an important step change in how researchers think about cleaning. Using this model may mean that the production of the chemical entity is not required to determine the cleaning agent required to remove it successfully from production vessels post manufacture. In addition the model may be used in the make stage of the process if there is a cleaning dilemma. If the tool developed in this research is to be used as part of FUSE then it is important to establish how this would fit in with corresponding Britest tools. Chapter 3 (section 3.5.2) discussed Britest tools and those which would be useful in providing information on selection of cleaning agents. It is considered that the tool developed in this research project should be used alongside Britest core methodologies, although it could be used as a standalone tool. Although there is often no specific order in which to use the Britest tools discussed, it is thought that methodologies could be considered to help address cleaning challenges in the following order (figure 6-2).

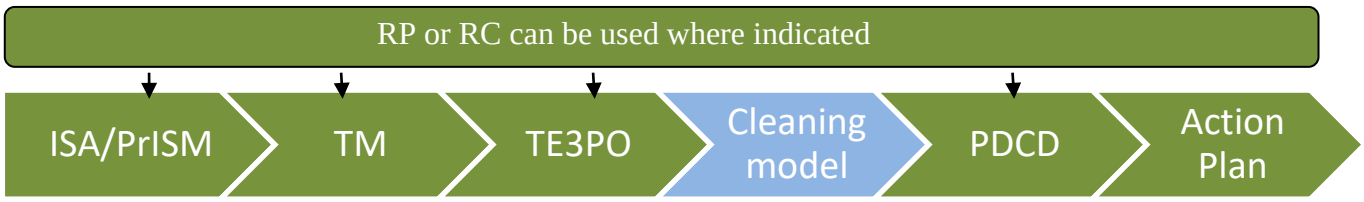


Figure 6-2 An example of a FUSE Roadmap.

Initially, it is important to define what the cleaning challenge concerns and identify what is not known. In order to achieve this ISA (used to define a problem and bring focus to it), or PrISM can be used. PrISM helps to summarise the process and determine what is happening at a stage in a manufacturing or indeed the cleaning process. This could be used to determine where residue is forming during processing. The use of Rich Pictures (RP) or Rich Cartoons (RC) may help identify the cause of this. Transformation Maps are an important Britest tool, which should be used in conjunction with the cleaning model. This is because it will help to identify desired and undesired processes when manufacturing a product. Identification of these processes may identify side reactions and intermediate products, which can be used in the cleaning model to determine their cleanability from process equipment (Chapter 3, section 3.5.2).

TE3PO provides information on physical processing and transformations. It allows formation of records and analysis of the data found. Chemical processes can be complex and this gives a logical approach to thinking through those processes, with regard to the entities that are present in the process stage, and what the physical properties might be. This tool may be able to help identify contaminants which cause residues in vessels (figure 3-16, Chapter 3), and may also be able to determine practical methods of contamination removal during this project.

The cleaning model could be used at this stage of FUSE. The cleaning model can be used with information provided from the previous Britest tools. The tool requires known structural and functional groups, which will have been identified by PrISM and TM. The information relating to the functional groups and structural features can be used in two ways. Firstly, the information can be fed into the original database and PCA can be carried out to determine the position on a score plot of the product, side reactant or intermediate. The product location on a score plot would help determine the best cleaning agent to use to remove it from vessels post manufacture. Secondly, it is also considered that the chemical and functional groups composing the product may be checked against a list of identified functional and structural features identified during this research as being found in products successfully cleaned from vessels (figure 5-21 reproduced as table 6-1).

Identified cleaning agents/ method	Chemical functional groups and structural features in identified in products cleaned from vessels post product manufacture
Water	Na+ Association, Hydrozone, Phosphate, Phosphonate, Tertiary alcohol association, Secondary alcohol, Ketone, Aryl halide, Steroid
Methanol	Tertiary alcohol structure, Vinyl alcohol, Oxime group, Macrolide, Tertiary amine, Tertiary alcohol, Ketone, Primary amide, Tetracycline, Secondary alcohol, Ester, Oxime, Ether, Erythromycin derivative
Methanol 1% HCL	Macrolide, Tertiary alcohol structure, Vinyl alcohol ,Tertiary amine, Secondary alcohol, Ketone, Ester, Ether, Primary amide, Tetracycline
DMF	Contain a mix of functional groups and identifying features Phenyl Ring, Primary amine, Secondary amine, Tertiary alcohol, Carboxylic acid, Aromatic enamine , Secondary alcohol, Secondary amide, Secondary amide, Primary amide, Ether, Carbamate, N-heterocyclic, Alkene, Alkyl >5 carbons, , Ketone, Oxazolidonone, Tertiary amide, Guanidine, Water, O-heterocyclic, Aryl halide, Sulfonamide, Macrocyclic, Primary alcohol, Tertiary amine, Carbamate, Urea, Barbitute, Thioester, Phenol, Long alkyl, Thioether, Nitro, O-heterocyclic, Sulfonamide, Vinyl alcohol, Phenyl ring, Ketone.
Acetone or DMF	No significant functional group identified. Do contain some common features Secondary alcohol, Ketone, Ester, Steroid, Alkyl halide, Water, Tertiary alcohol, Fluorine, Thioester, Ether

Table 6-1 Variables associated with products and the cleaning agents used to remove them from process equipment post manufacture. The black writing indicates features found in the products. Blue writing indicates common features identified in the analysis but not found in the products.

Figure 6-1 may be used as a quick check to determine potential cleaning agents by composition of chemical groups. Once these are known a cleaning agent may be selected. As this suite of tools considers Whole Process Understanding (WPU), it is important to have a tool in the suite which is able to help identify specific engineering challenges or materials, which may make cleaning from particular vessels difficult. The tool, which has been designed for this purpose is PDCD, an adaption of PDD. The PDCD described in Chapter 3 (figure 3-17) can indicate the age, material and staining of the vessel, in conjunction with a Rich Picture (RP). It can be used to show complex vessel geometry, which is often difficult to clean. It is considered that the choice of a cleaning agent cannot be made without using this diagram as cleaning needs to be carried out using a holistic approach. This requires taking into account the fundamental engineering challenges as well as the scientific understanding.

In addition to the FUSE roadmap (figure 6-2) additional tools RP and RC may be used to help target specific problems and focus on issues surrounding the vessel which requires cleaning, or the manufacturing process that is taking place specifically in one piece of equipment.

Once FUSE has been applied to a challenge an important part of the roadmap is to take action to make changes, which will increase the effectiveness of cleaning and aim to ensure that the next clean carried out post manufacturing is carried out RFT.

An additional tool, which can be used to help industrialists, is Duty Definition and Equipment Specification (DuDEs). DuDEs is used for process equipment decision making and may also be used to consider the selection criteria of new cleaning equipment. This may be carried out if specific cleaning equipment is required to clean a vessel identified in PDCD analysis.

Britest's tools and methodologies are very adaptable and this makes them suitable for the identification of many industrial challenges and their solutions. The aim of this research was to create a tool to help understand the fundamental science behind cleaning. In order to determine whether the cleaning model developed is useful to industrialists, it is important to use it to carry out case studies and test the theory behind the cleaning model. The case studies will be discussed in the next section 6.3.

6.3 Case Study Introduction

The suite of tools designed to help industrialists understand the science behind cleaning was discussed in section 6.2. In this section it was important to consider if the model developed during this research, based on understanding the fundamental science behind cleaning, is effective. In order to achieve this, two case studies were carried out. The first case study used information provided by company C. The second case study was carried out using information provided by company B. PCA analysis of data from both companies was carried out at the same time. Therefore there is only one set of PCA results. The case studies, the results obtained from the analysis and the conclusions drawn from the results are given in the following sections 6.4, 6.5 and section 6.6.

6.4 Case Study 1 Company C

The first case study involved information provided by company C. Company C is a large multinational company, which produces pharmaceutical products. Information provided by company C composed of functional groups and structural information for one product and 4 intermediate products for the same process. Company C were unable to provide information relating to physicochemical properties for any of the products they gave for the case study. This indicates that the information was difficult to obtain. It is thought that this was especially true of intermediate and side products. It was considered that using intermediate products was a good way to determine how the tool could be used to help decide which cleaning agent to use to clean vessels post manufacture. Industrialists often choose a cleaning agent based on

the final products solubility or a problem intermediate which is the most difficult to remove from a vessel. The information provided by company C did not indicate which of the chemicals was the most difficult to clean out of vessels. The company also gave no indication of the type of vessels which are involved in processing. The information provided by company C is indicated below (table 6-2). Table 6-2 shows the chemical and structural information, which was provided for each chemical labelled P1 to P5.

Table 6-2

Product name	P1	P2	P3	P4	P5
Identified position in process*	Intermediate product	Intermediate product	Intermediate product	Intermediate product	Final Product
Functional groups identified	Tertiary amine, primary amine, Fluorine atom, Alkene group, Alkyl group greater than 5 carbons	Tertiary amine, aromatic/ enamine group, carboxylic acid, Fluorine atom, Alkyl group greater than 5 carbons	Tertiary amine, Primary alcohol (OH) group, Fluorine atom, Alkyl group greater than 5 carbons	Tertiary amine, Ether group, Fluorine atom, Alkyl group greater than 5 carbons, Other feature (not identified)	Secondary amine, Fluorine atom, Alkyl greater than 5 carbons
Structural feature identified	Phenyl ring			Phenyl ring	2 Phenyl rings
cleaning agent used	Methanol Unsuccessfully used	Methanol Unsuccessfully used	Toluene Successfully used	Methanol Unsuccessfully used	Methanol (highly soluble in 50% solutions but still a cleaning challenge)
Cleaning agent suggested using information	DMF	DMF	DMF	DMF	DMF

Product name	P1	P2	P3	P4	P5
in table 6-1					
Cleaning agent suggested using PCA analysis and position on score plot	Acetone or DMF	DMF	DMF	DMF	Acetone or DMF

Table 6-2 contains the information that was provided by company C. *The intermediates and product in this table have been identified in one process.

Table 6-2 also shows the cleaning agents which company C currently uses to try and remove the listed chemicals. The cleaning agent used for most intermediate product removal is methanol. Methanol is a volatile and flammable liquid with a flash point of 52°c (11°c). It is classed as hazardous waste and it is harmful to aquatic life in low concentrations. The recommended method of disposal is burning.

Intermediate product P3 is removed from vessels by rinsing with toluene. Toluene may be a teratogen in humans. It is extremely flammable with a flash point of 40°F (4°c). Hazard classifications given to this solvent indicate that it is detrimental to health (GHSO7, GHS08). This would make it difficult to clean with this solvent especially in open process vessels. Toluene is described as hazardous waste and it must be removed via a chemical waste disposal service. This can become expensive as discussed in chapter 3.

Analysis was carried out by comparing the chemical functional groups and structural features associated with known cleaning agents found in the analysis (table 6-1). After examination of the information available for each case study chemical against the information in table 6-1, it is possible to say that DMF might be used to clean vessels containing residues of the chemicals listed by company C. This solvent has a higher flash point (136°F or 57.77°c) than the other two solvents currently used. This means it may be easier to use. In addition it is not considered as harmful to human health as toluene or methanol. Disposal of DMF requires a chemical waste disposal service, but if cleaning is carried out right first time the levels of solvent for disposal post cleaning may significantly reduce.

In addition to looking at the information in table 6-2 the new information provided by company C was normalised and analysed by PCA. This gave the results in section 6.4.1 (figures 6-2, 6-3 and 6-4). As the PCA analysis was carried out in conjunction with the data for company B, it was first important to discuss the information given for case study by company B for analysis before the results of the PCA are discussed. The case study for Company B will be discussed in section 6.4.

6.5 Case Study Two Company B

Company B is a large pharmaceutical and agrochemical manufacturing and contract manufacturing organisation. The information provided by Company B is shown in table 6-3. The information provided for the case study included TM, which was helpful in understanding the chemicals presented for the case study. However, because of confidentiality these cannot be reproduced in this research.

Product name	P6	P7	P8	P9	P10	P11
Identified position in process	Undefined	Undefined	Undefined	Undefined	Undefined	Undefined
Functional groups identified	2 Ketone groups, 1 Ether group, 1 Alkyl greater than 5 carbon,	1 other organometallic group	Unknown numbers of primary alcohol, ester, Alkyl greater than 5 carbon groups,	Unknown numbers of primary alcohol, carboxylic acid, ester, Alkyl greater than 5 carbon groups,	1 Alkyl greater than 5 carbon group	1 Primary amine, 1 secondary amine (both unknown polymers), 1 ester, 1 primary amide and 1 secondary amide.
Structural feature identified	1 O-heterocyclic group, 1 Long alkyl group	Other	Unknown numbers of Long alkyl group	Unknown numbers of Long alkyl group	1 Phenyl ring, 1 Long Alkyl group	An unknown number of N-heterocyclic features and an unknown

Product name	P6	P7	P8	P9	P10	P11
						number of Phenyl rings
cleaning agent used	Unknown but currently using water	Water	Polyisobutylene, caustic and water	Caustic soluble but cleaning as yet undefined	85% Phosphoric acid in water	Soluble in water
Cleaning agent suggested using information in table 6-1	DMF	Insufficient information	DMF	DMF	DMF	DMF
Cleaning agent suggested using PCA analysis and position on score plot	DMF	DMF	Acetone or DMF	DMF	DMF	DMF

Table 6-3 Information provided for case study by Company B.

Company B provided a lot of information about the chemicals it supplied for the case study. Table 6-3 shows some of this information. Each chemical was labelled sequentially following on from the chemicals listed in table 6-2 for clarity. The chemicals supplied were classified as undefined. This means they were not considered products or intermediates. There is not a lot of information supplied about any of the chemicals compositions, which means that the data that is usable in the PCA analysis is limited. This is not a surprise. The survey data in chapter 3 indicated that industrialists do not fully understand their processes and often do not know the composition of most intermediate chemicals, which remain undefined. The complexity of

manufacturing chemicals is appreciated and in these cases a TM or TE3PO analysis may have helped define the chemicals and increase understanding of their composition.

P6 was defined as insoluble, but water was being used to remove it from vessels post manufacture, as the chemical hydrolyses in water. Company B do not know how to remove this chemical from process vessels. Physicochemical data provided by Company B for this chemical was the density (0.96g/cm^3). P7 was not a challenging product to clean from vessels and water was used to remove it from vessels post manufacture. There was only limited information provided about this chemical, other than it contains an organometallic group, which is defined as other group in the PCA. Some physicochemical properties were provided for this chemical. These were values for the flash point (17°c), Density (1.013g/cm^3), Vapour pressure (173 bar @ 20°c), and the Boiling point which was 338-342K. As the amount of physicochemical data is limited it is not possible to use this information in the analysis. P8 was a chemical, which Company B found difficult to remove from vessels post manufacture. Polyisobutylene (PIB) was used to remove this chemical from vessels (PIB is used as an additive in engine fuel to prevent soot, sludge and other deposits from leaving residues on surfaces). P9 was an undefined chemical, which was a challenge to clean from vessels. Company B did not know how to remove this from equipment but did know that it was soluble in caustic. Physicochemical properties for this chemical were limited but the melting point was provided (<253K), the flash point (>24°c) and the density (0.98 @ 20°c). Company B provided information on the undefined chemical P10, which was challenging to remove from their equipment. They removed this chemical with 85% Phosphoric acid in water. They were able to provide very limited physicochemical data for this chemical, which was the flash point (191°c) and the melting point (393K). This chemical was insoluble in water and soluble in alcohol. The final chemical provided for the case study was P11, which was not a challenge to clean from vessels as it was soluble in water. It was cleaned from vessels using 85% phosphoric acid in water. The flash point (>100°c), the boiling point (100°c) and the density (1.15@ 20°c) of the chemical were known. Comparisons between the information provided in table 6-1 and 6-3 indicated in some cases there was not a lot of information provided to determine which cleaning agent could be used for chemical P7. The other chemicals P6 and P8 to P11 were all determined to be cleaned from vessels with DMF. This is probably due to a lack of information. The cluster of products which have been determined to be cleaned by DMF is large and therefore there are many functional groups associated with this cleaning agent. The 6 chemicals provided by company B for this case study were all undefined and a lot of the information provided was uncertain. This made the analysis difficult. A number of

functional groups (given in table 6-3) were listed as present, but the number of the functional groups was unknown. For the purposes of PCA this made analysis difficult as the number of functional groups of one type affects the analysis. Unavailable data meant that one functional group was recorded against the type in the PCA analysis where there could have been more. PCA of the information provided by Company B was carried out with the data provided by Company C in section 6.5.

6.6 PCA Analysis of the case study data for company B and company C

The case study information for both company B and C was analysed by PCA. The result and a discussion of this analysis are provided in this section. Figure 6-2 shows the scree plot produced from PCA analysis of the original data set and with the addition of the new case study data. All data was normalised.

6.6.1 Scree Plot analysis of the original data and the case study data

Analysis of the data began by examining the scree plot (figure 6-2)

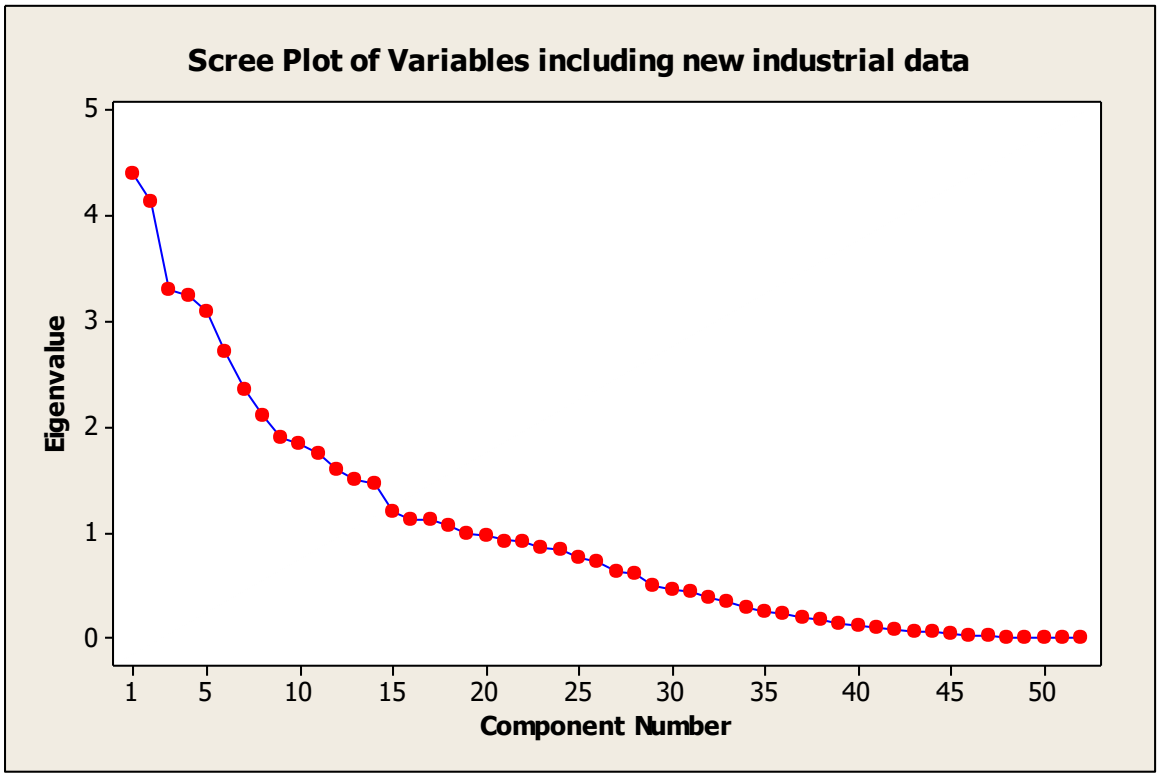


Figure 6-3 Scree plot from PCA analysis including data obtained from industrial case studies for both company C and company B.

Figure 6-3 indicates that there are potentially two elbow points in the data. The first point occurs at the third principal component. Data up to this point on the scree plot accounts for

22.8% variation in the data set. The second elbow point in the data occurs at the 14th principal component. The data on the scree plot, up to and including this principal component, accounts for 68.2% of the variation in the data. Therefore it is the first 14 principal components which add the most variation in the data set. Analysis of the first 14 principal components indicated the variables of interest (table 6-4).

Variable of Interest	**Principal Component number**	**Variable of Interest**	**Principal Component number**
Primary amine	C9, C13, C14	Phosphonate	C2
Secondary amine	C2, C3, C6, C12, C13	Hydrozone	C2
Tertiary amine	C3, C5, C10	Other	C2
Aromatic / enamine	C4, C5, C9	Phosphate	C2, C3, C5
Primary alcohol	C3, C10, C12	Carbamate	C12
Secondary alcohol	C1, C3, C4, C10	Nitro	C4, C5, C6, C9, C10
Tertiary alcohol	C1	Nitrate	C7, C9, C12
Vinyl alcohol	C4, C5	Steroid	C2, C6, C9, C11
Phenol	C6, C8, C12, C13, C14	Hormone	C8, C11, C12, C13,C14
Carboxylic	C3	O-heterocyclic	C4, C14
Ketone	C2, C4, C6	N-heterocyclic	C4, C7, C8, C9, C10
Thioester	C6, C10, C12	S-heterocyclic	C4, C5, C6, C10
Oxime	C1, C4, C13	Long alkyl	C10,C12, C13, C14
Oxazolidinone	C14	Phenyl ring	C6, C8, C9, C11

Variable of Interest	Principal Component number	Variable of Interest	Principal Component number
Urea	C8, C11	Erythromycin derivative	C1, C4, C13
Guanidine	C9, C12, C13	Tetracycline	C1, C4, C13
Ether	C1, C9, C12, C13, C14	Macrocyclic	C4, C5
Sulfonamide	C12, C14	Macrolide	C1, C13
Sulfone	C11, C12, C13, C14	Benzodiazepine	C10, C12
N-oxide	C7, C9, C12	Barbiturate	C8, C11
Thioether	C4, C5, C6	Water	C5, C9
Fluorine	C2, C6, C10	Ethanol	C4, C5, C6
Pyridine	C2	HCL	C4, C5
Alkyl halide	C6, C9, C10, C11	Na+	C2, C3
Aryl halide	C6, C8, C11, C13	Gd3+	C3, C4
Alkene	C9, C11, C12, C13, C14		
Alkyl greater than 5 Carbons	C9, C11, C12, C13, C14		

Table 6-4 Principal components identified in the scree plot as contributing to the variability in the data set. Variables which added the most variability in the first 3 principal components are highlighted in red.

Analysis of the scree plot indicated that every variable added to the variability of the data set in the first 14 principal components. The data showed the variables which added to variability in the first 3 principal components. This was where the most variation in the data set was found. These were highlighted in red in table 6-4. These variable groups and structural features were listed as Phosphonate, Hydrozone, Other, Phosphate, Steroid, Erythromycin derivative, Tetracycline, Macrolide, Na+, Secondary amine, Tertiary amine, Primary alcohol, Tertiary amine, Primary alcohol, Secondary alcohol, Tertiary alcohol, Carboxylic acid groups, Ketone, Ether, Fluorine and Pyridine. This list was compared to the list of functional groups and structural features, which were found to be of interest in the analysis of database 1 (Chapter 5, table 5-3). The following similarities were found. Both analyses identified the variables primary and secondary amine groups, phosphonate groups, phosphate groups and carboxylic acid groups. Analysis of the score plot form the PCA analysis was carried out next (section 6.5.2) in order to determine more information on the case study chemicals.

6.6.2 Score plot analysis of the original data and the case study data

The score plot was a useful tool in previous analysis of this research in chapter 5. It was able to show a link between data on cleaning agents and the products, which in turn gave information on shared functional and structural features. The score plot generated for analysis of the case studies is shown below (figure 6-3).

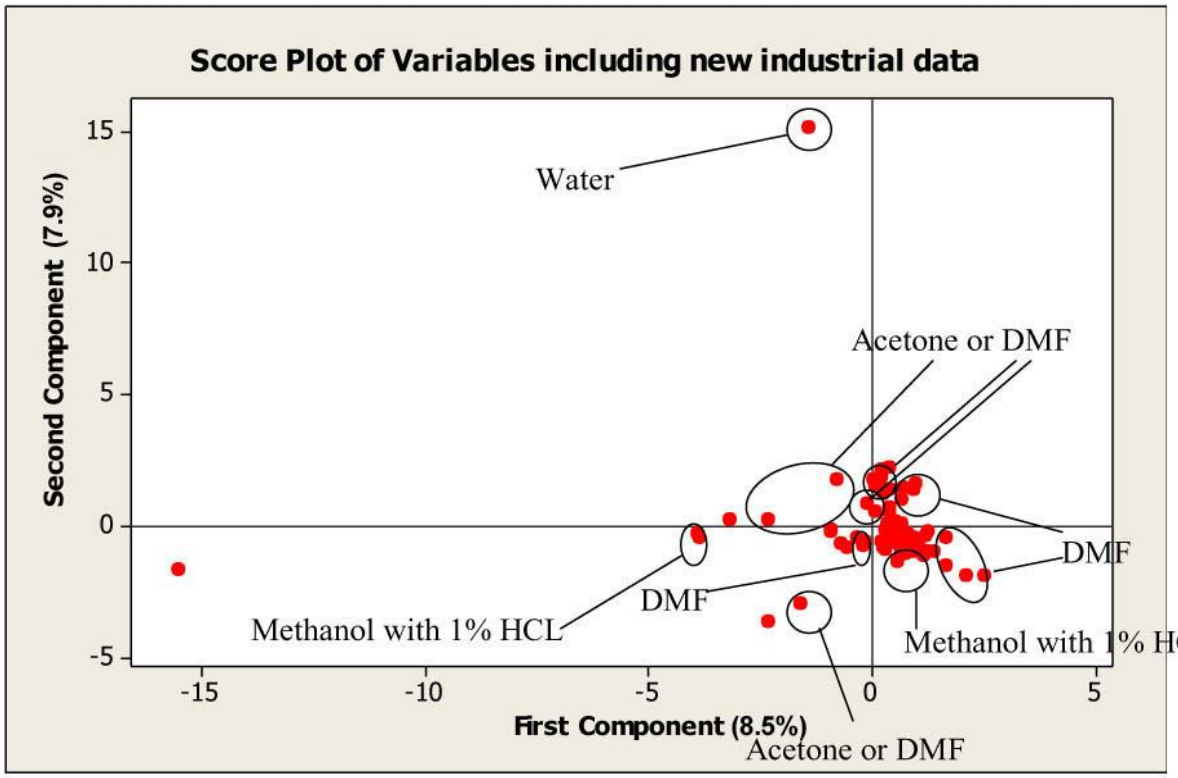


Figure 6-4 Score plot from PCA analysis including data obtained from industrial case studies for both company C and company B.

PCA analysis produced the score plot (figure 6-4). The figure shows a lot of data points located around the zero axes. There are some data points which are located distinctly away from others on the plot. Annotation of the score plot indicates some of the data points, which are associated with different cleaning agents, (as determined by analysis of database 1 in Chapter 5). The data was complex to analyse but it gave the following information on the case study chemicals. P1 was found on the score plot located next to chemicals, which had previously been identified as being associated with the cleaning agents Acetone or DM. P2, P3, P4 P8, P10 and P11 were not located on the score plot. P5 was associated with products, which are cleaned from vessels post manufacture by Acetone or DMF. These are very different from the cleaning agents, which are listed in table 6-3. The next section examines the information provided in the Loading plot in section 6.5.3.

6.6.3 Loading plot analysis of the original data and the case study data

The PCA gave three plots, which provided information to help identify patterns and links within the data. This is the third plot, which will be used to identify the variables that provide the most variation in the data set. The Loading plot is shown below (figure 6-5).

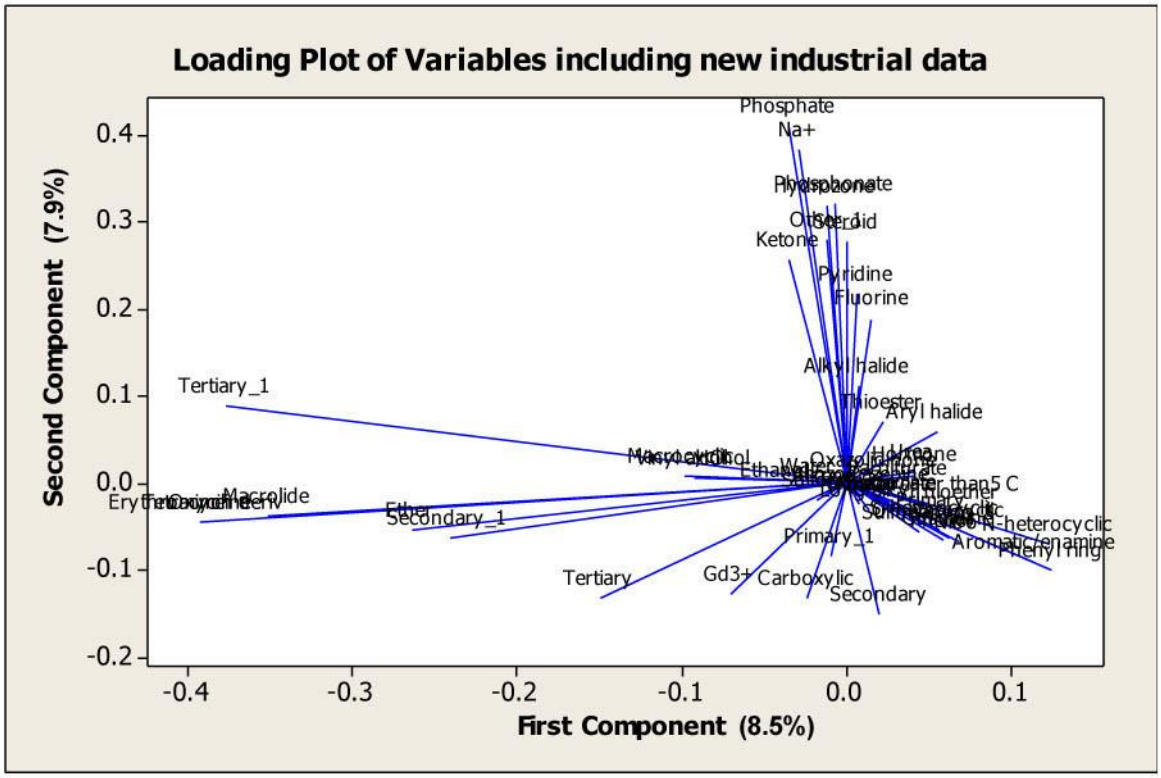


Figure 6-5 Loading plot from PCA analysis including data obtained from industrial case studies for both Company C and Company B.

Figure 6-5 indicates the variables of interest. Using the information in table 6-5 it is possible to see the variables, which provide the most variation in the data set. A lot of the variables are located around the zero axes. Those which add the most to the variability of the data set are those which are physically distinct from this area on the plot. The variables which are thought to add the most variation to the data set are Phosphonate, Hydrozone, Other, Phosphate, Steroid, Erythromycin derivative, Tetracycline, Macrolide, Na+, Secondary amine, Tertiary amine, Primary alcohol, Tertiary amine, Primary alcohol, Secondary alcohol, Tertiary alcohol, Carboxylic acid groups, Ketone, Ether, Fluorine and Pyridine. In addition the variables have clustered on the plot (figure 6-5). Some of the clusters include Tertiary alcohols, Oxime and features Tetracycline, Erythromycin derivative and Macrolide. It should be noted that although these variables show the most variability in the database, none of the chemicals in the case study contain these features. Variables which are considered to add the most to the variability in the data set are not found in the chemicals provided for the case study, with the exception of Ketone (P1), Primary alcohol P3 and Phenyl ring P10. In addition, variables provided by companies were not found within the scree plot to add considerably to the variability. This information indicates that it may be difficult to identify a cleaning agent for the chemicals in the case studies due to insufficient data. In order to understand the data better

it was appropriate to rerun the analysis using only the data, which was located around the zero axes. This will be carried out in section 6.5.4.

6.6.4 Analysis of the main data set located around the zero axes.

Analysis of the main data set located around the main axes (figure 6-3) was carried out to obtain a better understanding of the data. PCA was rerun on data used in section 6.5.2. Several products which lay outside of the data required were removed from the dataset. The products removed were Betamethasone disodium phosphate, Gopten, Halbetasol, Betamethasone acetate, Oxis, Meperidine, Fluticasone furoate, Epival, Clobetisol propionate and Plendil. The removal of these products was based on the physical location on the score plot in relation to the data of interest. PCA analysis was carried out and produced the following score plot (figure 6-5). Only the score plot is shown from this analysis because this is the plot which will show the relationship between the chemicals in the case study and the products used in the model development. It was considered that because of the products locations on figure 6-3, all of these products should have a relationship and that they may be considered to be linked with the cleaning agents DMF or Acetone.

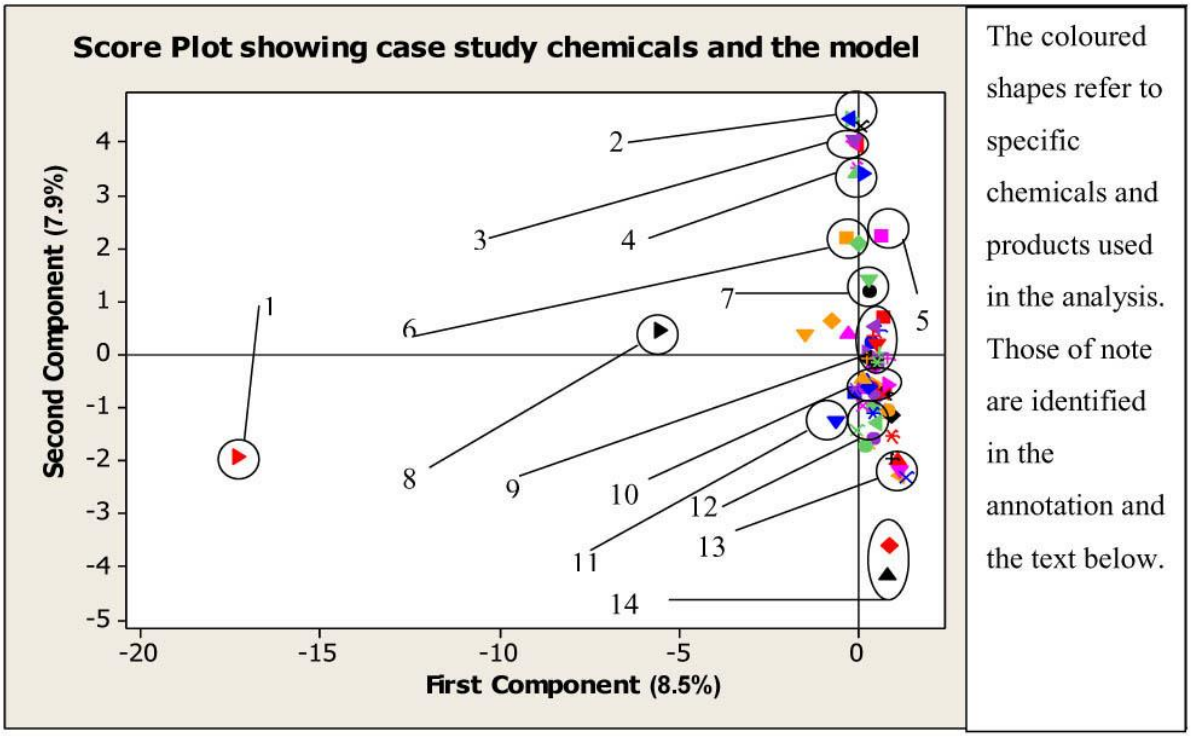


Figure 6-6 Score plot showing the relationship between the case study chemicals and the products used in the model. The coloured shapes refer to specific case study chemicals and products used in the original model. Those chemicals and products of note are identified by

the annotations. It seems as if the chemicals from the case studies are linked to the cleaning agents DMF or Acetone.

There were 14 groups and points of interest identified on figure 6-6. Each of these will be discussed below.

Groups or points of interest identified on figure 6-6 are described below -

1. This point refers to the chemical Selelamer, which was not located on the previous score plot used as the model (database 1score plot). The cleaning agent for this product remains unknown but it does not appear to be physically located near the rest of the data on this score plot.

2. The second group of interest contains the pharmaceutical products Hytrin, Nimbex and Nizatidine. All of these products were associated with different groups in the previous analysis of the data set prior to the addition of the case study data. Hytrin is associated with the cleaning agent DMF, Nizatidine was associated with group 5 and Nimbex was associated with group 12. It is not known why these products were clustered in this way in this analysis.

3. The third group contained the products Betamethasone dipropionate, Betamethasone dipropionate monohydrate and Cycloserine. All of these products have previously been associated with the cleaning agents DMF or Acetone.

4. The fourth group contained the chemicals Gabopentine, Doxycycline monohydrate and Furosemide. Both Gabopentine and Furosemide have been associated with the cleaning agent DMF.

5. The fifth point of interest was the product Metronazole. This product was not associated with any cleaning agent in the model and it was not identified in the analysis (Chapter 5, table 5-2).

6. Both products identified in this group (Citanest and Androgel) were associated with the cleaning agent DMF in the previous analysis.

7. The two products identified in this group were Deflox, associated with the cleaning agent DMF, and Advicor which was associated with group 10 in previous analysis (Chapter 5, table 5-2).

8. Conolip was the only product found at this point. This was not identified as being associated with any cleaning agent in the research. This was due to limited availability of data.

9. This group of products was found to include Aluvia, (previously associated with group 12 in the research (Chapter 5, table 5-2)), and Isradipine, (which was previously associated with group 9). The rest of the group contained the case study chemicals P1, P9, P10 and P8.

10. This group of products included Warfarin, Iodixamol, Atenonol and Brofen. All of these products were associated with the cleaning agent DMF in the research (Chapter 5, table 5-2). The following case study chemicals were found in this group. This included P2, P4, P5, P6, P7 and P11, although these chemicals had different cleaning agents reported by Company C and Company B. The combination of functional groups and structural features, which are associated with this group are diverse (table 6-1). It is considered that when there is limited information on a product or a chemical it is difficult to predict a cleaning agent.

11. Point eleven was identified as Sumatriptan Base, which was not previously identified in the analysis.

12. This group contained the case study chemical P3 which was the only known chemical or product in the analysis to be cleaned from vessels using Toluene. This was clustered with products which were associated with the cleaning agent DMF. These were Severane, Metaprobamate, Quinapril and one product (Mometasone furoate anhydrous), which was previously associated with the cleaning agents Acetone of DMF.

13. This group contained a number of products which were previously identified in the research to be associated with the cleaning agents Acetone or DMF. These were Marcaine, Mometasone furoate monohydrate, and two products whose link to cleaning agents is unknown, (previously associated with group 7 in the analysis (table 5-2 Chapter 5)). These were Gadopentetate dimeglumine and Gadopentetate monomeglumine.

14. The final group of interest contained the products Salmeterol xinafoate (associated with the cleaning agent DMF in this research) and Folic acid. Folic acid had not been associated with any cleaning agent in this research.

6.6.5 Conclusion

The products and chemicals identified on the score plot (figure 6-4) indicate that there are relationships in the data and other factors which have not yet been identified, and that affect

the results. It is considered that more information needs to be gathered on cleaning agents which can be used to populate the model. The information about the chemicals, which was provided by Company C and Company B for the case study, seem to be linked with the cleaning agent DMF. Considering the information provided by both companies shown in tables 6-2 and 6-3 did not mention the cleaning agent DMF, there must be other factors which are influencing the data. Factors which could be affecting the results could include that not much data was available on the case study chemicals. In fact, the data which was included in the PCA analysis was not a true reflection of the composition of the undefined chemicals, as the true composition of many of the chemicals was unknown (table 6- 3, (P8, P9 and P11)). This was not sufficient and the lack of information may have had an effect on the analysis. Another factor influencing the results may be the type of chemical functional groups, which are found in the case study chemicals. In all cases the functional groups and structural features were not strongly identified in the data. It is believed that the combination of the functional groups, not just the presence of the functional groups, indicate the cleaning agent which should be used to clean products from vessels post manufacturing. In addition, it should be noted that the model was based on information provided by one company. It may therefore be considered that different cleaning agents are cleaned from vessels using different cleaning agents with different degrees of success. This information is not currently available but it is thought that additional input from companies would greatly increase understanding of the model, and of the fundamental connection between the science behind cleaning and the selection of a cleaning agent. Finally, the model was constructed using data provided for products and not intermediates or side products, which are undetermined. The product information is better defined than the chemicals provided in the case study. Therefore it may be possible to conclude that data of this type requires its own model, which is not based on product data. This will be investigated in section 6.5.5.

6.6.6 PCA analysis of the Case study data

The information provided by Company C and Company B was normalised and analysed by PCA on its own to determine if any links in the data could be determined. It was considered that there was insufficient data to carry out this analysis but for completeness the analysis was performed. Figures 6-6, 6-7 and 6-8 show the PCA results. Figure 6-6 shows the results of the analysis via the scree plot.

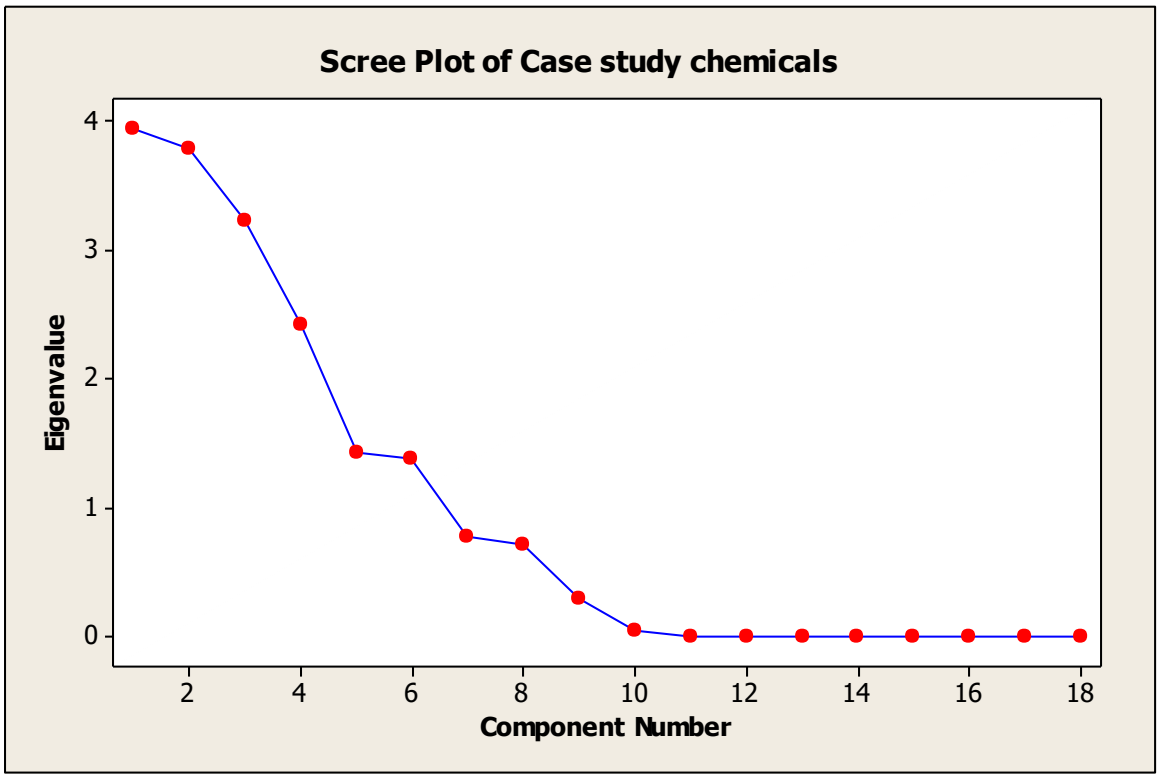


Figure 6-7 Scree plot from PCA analysis performed only on the case study chemicals.

The scree plot shows that a high proportion of the variables are represented as showing variability in the data. This is because the amount of data used in the analysis is limited. There are 10 variables which appear to contain 100% variation in the data set. There appear to be two ‘elbow points’ in the data set. The first point appears to be at component 5, which accounts for 82.3% of the variation in the data. The second ‘elbow point’ appears to be at component 8. Component 8 accounts for 98.2% variation in the data set. The first component in the data accounts for 21.9% of variation in the data. The last 9 principal components account for none in the variation in the data. The variable which appears to account for the most variability in the data is Alkyl >5 carbons. This variable appears to add to the variation in the first 4 principal components. Primary Amine functional groups and Fluorine also add the greatest variation in the group.

In order to examine the data further it is necessary to look at the information provided on the Score plot. This was carried out with figure 6-8.

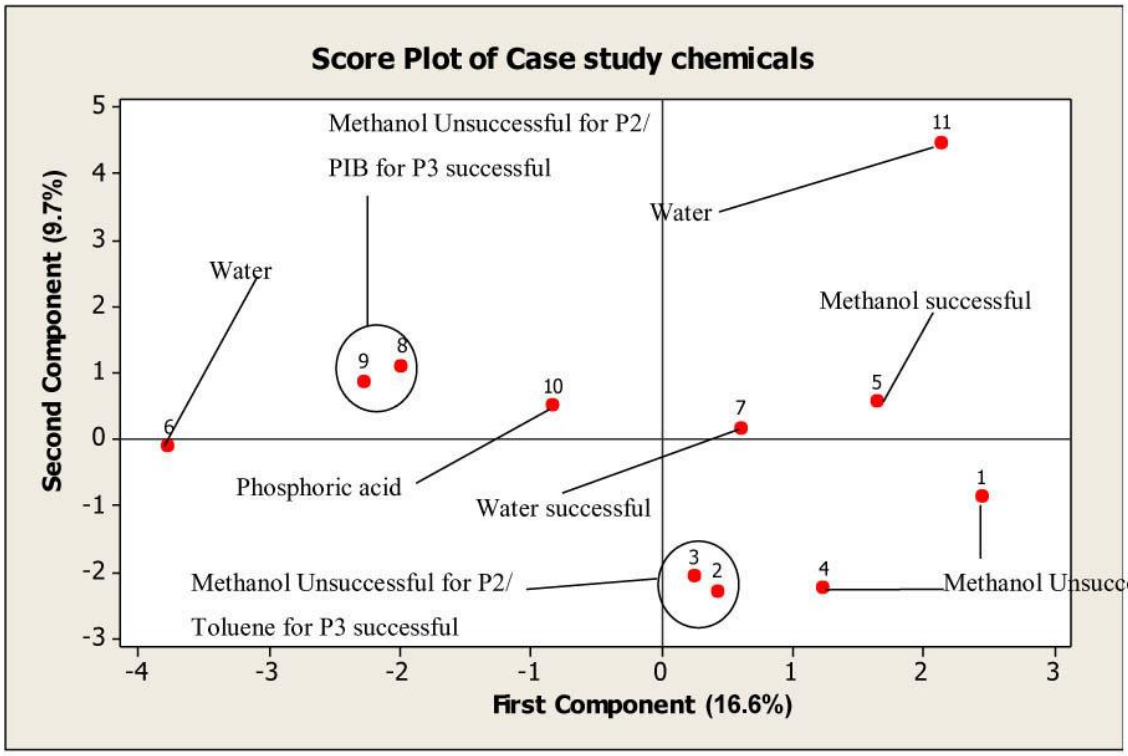


Figure 6-8 Score plot from PCA analysis performed only on the case study chemicals. The numbers refer to the reference number of the chemical in the study. Annotations indicate groupings of chemicals, and also where chemicals are associated with particular cleaning agents, according to Company B and Company C.

Figure 6-8 indicates the location of the chemicals on the score plot. The score plot indicates that the chemicals were widely distributed on the plot. There were a few notable groupings, which are shown annotated on the plot (on figure 6-8). There are two groupings; the first contains chemicals P2 and P3. The chemical P2 is unsuccessfully cleaned from vessels using methanol but P3 is successfully cleaned from vessels using Toluene. It may be considered that Toluene might be a good choice of cleaning agent to use to try and clean P2 from vessels post manufacture. This choice is recommended without an understanding of the vessels and the materials they are made from. This information should be identified using FUSE to ensure appropriate cleaning agents are chosen.

The second grouping shown on figure 6-8 is P8 and P9. Although both chemicals are cleaned from vessels with potentially different cleaning agents by Company B, both of these chemicals have unknown or undefined numbers of functional groups (table 6-3). It is not possible to easily determine the cleaning agent used for these chemicals. Both cleaning agents

PIB and caustic could be considered choices as both are used to some degree to remove these chemicals. The cleaning agent for P9 is not yet known and PIB may be a good choice of cleaning agent.

Points on figure 6-8 were annotated to indicate the cleaning agent used. This showed that chemicals cleaned from vessels using the same cleaning agents had not clustered. It is considered that there is not enough information on the chemicals or chemical variables in the database to allow this. In order to complete the analysis the loadings plot is investigated, figure 6-9.

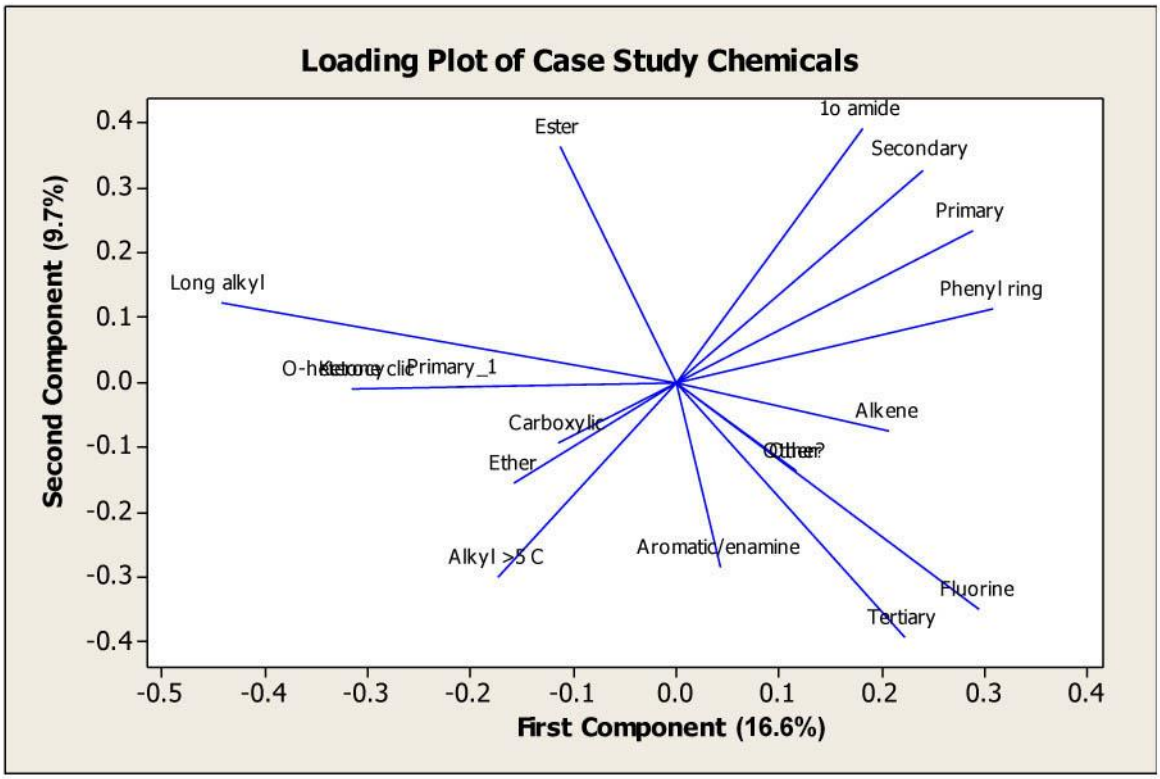


Figure 6-9 Loading plot from PCA analysis performed only on the case study chemicals.

It is considered that there is not enough information presented in figure 6-9 to be able to derive significant conclusions. The plot indicates that the variables represented add to the variation of the data set.

Conclusions have been drawn from the analysis of the chemicals presented in the case studies by Company B and Company C. These will be discussed in section 6.6

6.7 Chapter Summary

Analysis of the case study chemicals has not been able to definitively determine which cleaning agents should be used to remove them from vessels post manufacture. The

information was initially processed by comparing it with data (table 6-1). This indicated that although there was not a lot of data presented for each chemical, most of the information was thought to align with the group of chemicals which were cleaned from vessels with DMF. The lack of physicochemical information meant that the data selected for the model constructed from database 1 information (Chapter 5) was appropriate. It is considered that physicochemical data on products is difficult to obtain and therefore the provision of little data of this nature for the case studies was not surprising. PCA analysis was carried out. This showed that all of the case study chemicals seemed to locate on the score plot next to products which were linked to the cleaning agents DMF or acetone. Further PCA analysis of the products and the case study chemicals linked to the cleaning agents DMF and acetone, indicated that most of the chemicals had clustered together. The chemicals were all considered linked to the cleaning agent DMF when this analysis was examined. In order to establish if the data would yield any more information, clustering or linkages, it was normalised and PCA was again carried out. This time only the case study data was analysed. This showed that the data did not cluster according to known cleaning agents. The data showed two clusters of information amid distinct points. The two clusters mean may that a similar cleaning agent could be tried for each chemical. The main reason for the data failing to cluster was probably the limited amount of data used in the analysis. This does not result in an extraction of meaningful results with PCA.

It should be noted that the provision of cleaning agents for the case studies and the data which was used in the original model may not be complete. It is known that many companies manufacture the same chemicals, which is why a lot of cleaning data is confidential and is difficult to obtain. A cleaning agent which works RFT is a competitive advantage. Companies who manufacture the same chemical may successfully clean the residues from vessels using different approaches which are both successful. It would be of benefit to obtain this data and use this to inform future case studies, and increase the size of the cleaning knowledge database/tool.

Chapter 7. Conclusions

7.1 Introduction

This chapter concludes the thesis by discussing the answer to the initial research question: RQ1: What would be the best way to increase the fundamental understanding of the science behind cleaning linked to solvent and cleaning agent use?

This chapter also discusses the thesis contributions, conclusions are summarised, and future work is considered.

7.2 Discussion

The main aim of this thesis was to suggest the best way to increase the fundamental understanding of the science behind cleaning linked to solvent and cleaning agent use. This was to be achieved by creating, modifying or suggesting existing Britest tools. Chapter 3 identified through site visits and questionnaires that there was a need for a better fundamental scientific understanding of industrial plant cleaning. It was believed that pharmaceutical plant cleaning was often neglected and not considered part of processing. The main contaminants forming residues in vessels were not fully understood and the cleaning agents selected to remove them was based on solubility rather than functional group or structural properties of the pharmaceuticals. Cleaning processes and protocols were not optimised and cleaning was therefore not often carried out RFT. This had a detrimental effect on processing schedules and resources.

In addition, site visits conducted with Britest members identified that any tool developed needed to show cleaning challenges associated with the engineering aspects of cleaning pharmaceutical plant. These challenges range from difficult materials to clean, hard to reach places because of plant geometry and the age of plant vessels. It was considered that all of these aspects would be too complex to incorporate into one tool therefore a suite of tools was required. The Britest tools were examined and discussed with an aim to finding complementary tools and methodologies which could be used with the new tool developed during this research. The objective was to help achieve WPU and ideally, any tools developed should potentially eliminate the need for the manufacture of a pharmaceutical product in order to test its solubility.

The cost associated with inadequate cleaning was considered and Benson's ZEAL tool (developed for the food and drink industries for aqueous cleaning) was used to understand

cleaning costs. However, using Zeal was not effective for pharmaceutical cleaning as industrialists were not able to provide monetary values for their operations. In addition, ZEAL was not designed to consider the costs of waste disposal. Aqueous cleaning waste in the food and drink industries is generally cheaper and easier to dispose of. In the pharmaceutical sector chemical waste can be a major contributing factor to the cost of cleaning.

Chapter 2 discussed the literature available on cleaning. The literature discussed indicated that a lot of information on pharmaceutical plant cleaning was not available in the public domain. There are a few reasons for this. Industry has not traditionally given any thought to cleaning and therefore has not considered gathering data, let alone publishing it. It is thought that a successful cleaning regime gives any manufacturer a significant advantage over competitors and therefore the information is not shared if it is known. There are tools which can be utilised to increase the understanding of cleaning, such as the Britest tools, but prior to this research they have not been used for this purpose. The literature indicated that there was no published research on tools or methodologies which could be utilised to increase understanding of the science behind plant cleaning.

Chapter 4 discussed selecting and collating the data needed to understand the fundamental science behind cleaning. It was decided that the most fundamental data on pharmaceutical products was the structural and functional composition. This data was analysed by several methodologies until PCA was selected as the methodology to achieve the main aim of the research.

Chapter 5 discussed the use of PCA analysis to create a tool used to begin to understand the science behind cleaning. This was carried out using two databases of information on a number of pharmaceutical products using information in the public domain. One database contained information of functional groups and structural features of API's and the second contained information on the physicochemical properties of the same API's. Analysis indicated a number of functional groups, structural features and physicochemical properties clustered, or were shown to be linked in the data. This was shown on a score plot for each database. Known industrial cleaning agents for specific API was then linked to the information on the score plots. This showed that the cleaning agents were clustered around API's containing specific types of features and properties.

This information was collated into a table which was used to determine cleaning agent selection for case studies with Company B and Company C in Chapter 6. The case study information provided was different from the API data collated. The information obtained

concerned undefined chemicals. These were not defined as intermediate products or side products. The structural or physicochemical characterisation was not complete. It was difficult to analyse this information, as there was not much information available to analyse. Nevertheless, cleaning agent selection was carried out for each undefined chemical and in addition PCA analysis was carried out which confirmed the choice of cleaning agent selection with the limited data (with the cleaning agent selected using the collated table of information). Further PCA analysis was carried out using only the information provided by the two case studies. The analysis showed that the case study chemicals failed to cluster according to the case study companies selected cleaning agent. It was thought that the information provided for these case studies was neither precise nor sufficient enough to analyse in these circumstances. A Transformation Map of the process prior to carrying out PCA would have helped to determine more information about the chemical functional groups and structures.

7.3 Thesis Contributions

In summary the main contributions of this thesis are as follows -

- An extensive literature review identified and described methods used for cleaning in industry with a view to finding common cleaning methods in industry.
- A further literature review described current analytical methods used in industry to determine the cleanliness of equipment.
- Collated information on the regulatory documentation and regulations which are applicable to pharmaceutical plant cleaning.
- Structured questionnaires and site visits identified the industrial challenges associated with pharmaceutical plant cleaning.
- Established the level of understanding around plant cleaning in the pharmaceutical industry.
- Developed a tool which increases understanding of the fundamental science behind pharmaceutical plant cleaning.
- Modified existing Britest tools and identified others to create a suite of tools (FUSE) for use in pharmaceutical plant cleaning and the chemical industries.

7.4 Conclusions

In summary the main conclusions of this thesis are as follows -

- Pharmaceutical and chemical companies require a better scientific understanding of plant cleaning.

- A better scientific understanding of pharmaceutical plant cleaning can be achieved by obtaining information on specific functional groups and structural features of API's and carrying out PCA.
- PCA analysis can be used to determine clusters and patterns relating to the composition of chemicals. This information can then be correlated to known cleaning agents for specific chemicals. Known functional groups and structural features of these API can be used to identify other chemicals which could be cleaned from vessels using the same cleaning agent.
- A newly proposed methodology for the selection of cleaning agents and solvents for API's based on their functional groups and structural features.
- A method utilising physicochemical characteristics of API's was not successful because this information was not easily available for API's and therefore it was very difficult to obtain for side products, intermediate products and undetermined products.
- Cleaning agent selection was not achieved during analysis of case study chemicals which were described as undefined chemicals by the case study companies. It is thought that more precise information would have improved the analysis. Tools such as the Britest Transformation Maps, used correctly, would provide much of the relevant information on chemical structure to facilitate cleaning agent selection.
- It is considered that the model produced during this analysis for API's is not appropriate for side products, intermediates or undetermined chemicals. This is because the model is based on API data and the API chosen for the model were well characterised. Therefore, it is recommended that either a new model is constructed for side products, intermediates or undetermined chemicals or better characterisation of the chemicals is needed. Better characterisation will give better analysis of the chemicals using the original model.
- The new tool can be used in two ways for the selection of cleaning agents for API's. The first way is to determine functional and structural groups in the table which have been linked to cleaning agents and compare these with the features of the chosen API. The second method is to use the loadings that were identified from the original data to calculate new scores and determine where the chosen API is located in relation to API's with known cleaning agents.
- This database will be more useful to industry if it is further populated with more data. Specifically, it should be populated with API's with known cleaning agents. This will

help to determine if different companies use different cleaning agents to successfully remove the same product from vessels post manufacture.

7.5 Future Work

This section addresses particular areas where further research is required.

7.5.1 Future Case Studies

Britest Ltd intends to add this tool as part of the FUSE suite to their collection of tools and methodologies. The result of this will be the development of further case studies with the industrial members. These case studies will then be used in order to validate the findings of Chapter 5.

API's should be used for the case studies which would increase the data set used in the model. In addition, the opportunity to use FUSE as a suite of tools to help identify and resolve cleaning challenges in a case study on a site visit to a pharmaceutical company would validate the suite of tools.

7.5.2 Future Research Recommendations

It is recommended that the API model is further populated with more information to increase its robustness and effectiveness.

It is recommended that further data analysis (cluster analysis) is carried out on the databases to both complement the PCA and provide further information. Other software packages such as R, which includes algorithms such as AGNES (Agglomerative NESting) would allow further examination of the data. It is recommended that AGNES, which uses a bottom up clustering approach, or DIANA (Divisive ANAlysis) a top down approach to clustering, are used to determine if any further linkages or patterns are found in the data. The data used in this research relates to API's, their physicochemical properties and structural features. The analysis and initial conversations with industrialists has indicated that it may be possible for more than one cleaning agent to successfully remove an API from pharmaceutical equipment. The use of an algorithm which allows fuzzy clustering such as FANNY and not hard clustering which was used in this research may indicate dual cleaning agents to use for certain API.

It is considered that the PCA principal components could be further analysed in this research. This may indicate further links and patterns in the data which were not indicated in the first few principal components.

It is recommended that Britest Ltd members share cleaning information specific to commonly manufactured API's. This will help determine whether API's can successful be cleaned from vessels by more than one cleaning agent. This information would greatly increase the effectiveness of the existing model but it was not available during this research.

It is recommended that Britest Ltd obtain cleaning information for side products and intermediate products which have been better defined by the use of FUSE (using PrISM or TE3PO). This date could then be used to construct a model.

During analysis of the data, consideration was given to the creation of another database of API physicochemical and structural information from non Britest member companies or data randomly generated by the software. This was not carried out during this research. It was considered that using this approach it would be difficult to obtain cleaning data for any non Britest companies indicating either successful cleaning or unsuccessful cleaning. A random generated set of data would not be linked to any cleaning agents. Knowledge of cleaning agents is critical to begin to understand the key variables which indicate the use of cleaning agents for specific API. It is recommended that proceeding with either of these approaches should happen only after the addition of Britest member data to increase cleaning knowledge.

This research thesis focused on the challenges associated with cleaning pharmaceutical plant post manufacture. It is considered that the methodology used in this research may be useful in selecting final product formulations based on their functional and structural features and physicochemical characteristics. It is possible that these could be linked to known successful formulations which are well documented in the pharmaceutical literature. This may indicate formulations for new drug products and help to select new formulations for existing products.

Bibliography

ACD Labs (2016) LogD information. Available at http://www.acdlabs.com/products/percepta/predictors/logd/

Adamski, K, Bellinghausen, Voelkel, A (2008) Journal of Chromatography 1195 146-149

Aharoni, S.M (1992) Journal of Applied Poylm Science 45 813-817

Allen Associates 2016 (Accessed online January 2016) Available at http://www.allenhpe.co.uk/services/chemical-process-engineering-design.html

Al-Obeidani, S.K.S, Al-Hinai, H, Goosen, M.F.A, Sablani, S, Taniguchi, Y and Okamura, H. (2007) *Chemical cleaning of oil contaminated polyethylene hollow fiber microfiltration membranes*. Journal of Membrane Science 307 299 – 308.

Alsante, K.M and Hatajik, T.D et al (2001) *Isolation and Identification of process related impurities and degradation products from pharmaceutical drug candidates*. Part I American Pharmaceutical Review. (2001); 4(1):70-78

AMRI (2011) Information provided at site meeting Sept 2011.

Arslan, H, Günal, H, Güler, M, Cemek, B, Acir, N, (2013) *Assessment the soil properties affecting salinity and sodicity of bafra plain using multivariate statistical techniques*. Carpathian J. Earth Environ Sci 8(1):81–90.

AstraZeneca (2008) Cleaning Information. ISPE UK Affiliate Conference, Manchester 15 Nov 2008.

Augustin, W, Fuchs, T, Föste, H, Schöler, M, Majschak, J.P and Scholl, S (2010) *Pulsed Flow for enhanced cleaning in food processing*. Food and Bioproducts Processing 88 384 – 391.

Baldeschwieler, E.L, Troeller, W.J. Morgan, M.D (1935) *The Kauri Butanol Test for Solvent Power.* Ind. Eng. Chem. Anal. Ed. 7 (6) 374-377

Benson, R and Ahmad, M. (1999) *Benchmarking in the Process Industries*. Institution of Chemical Engineers. Rugby

Benson, S.W and Buss, J.H (1958) *Additivity Rules for the Estimation of Molecular Properties.* Thermodynamic Properties J. Chem.Phys. 29 (2) 546

Benson, S. W. (1968) *Thermo chemical Kinetics*, Wiley, New York.

Berridge, J (1995) *Impurities in drug substances and drug products: new approaches to quantification and qualification*. Journal of Pharmaceutical and Biomedical Analysis. 14 (1995) 7-12.

Bharathi, Ch, Prasad, Ch.S et al (2007) *Structural identification and characterisation of impurities in ceftizome sodium*. Journal of Pharmaceutical and Biomedical Analysis 43 (2007) 733- 740.

Bird, M.R. (1994). *Cleaning agent concentration and temperature optima in the removal of food based deposition*. Fouling and cleaning in food processing (p156-166). UK. Department of Chemical Engineering. University of Cambridge.

Bird, M.R and Fryer, P.J (1991). *An experimental study of the cleaning surfaces fouled by whey proteins*. Transactions of the Institution of chemical Engineers. C, 69 p13-21.

Bishop Y.M.M, Fienberg, S.E, Holland, P.W (1976) *Discrete Multivariate Analysis: Theory and Practice*. Cambridge: MIT Press.

Block Engineering (2012) *Noncontact, Real-Time Cleaning Verification for Pharmaceutical Manufacturing* White Paper/ Application notes. Accessed online January 2016. Available at at http://www.americanpharmaceuticalreview.com/1489-Application-Notes/116171-Noncontact-Real-Time-Cleaning-Verification-for-Pharmaceutical-Manufacturing/

Boly, M, Perlbarg, V, Marrelec, G, Schabus, M, Laureys, S, Doyon, J, Pélégrini-Issac, M, Maquet, P, Benali, H (2012) *Hierarchical clustering of brain activity during human non-rapid eye movement sleep*. Proceedings of the National Academy of Sciences of the United States of America, Vol.109 (15), page 5856-61.

Bott, T.R (1995) *Fouling of heat exchangers*. New York. Elsevier.

Britest. (2011) Information from Britest Training Documentation. Britest Ltd Training folder.

Britest (2011) Training course information, Britest Ltd.

Britest Ltd (2014). Britest Limited. Available at: http://www.britest.co.uk

Britest Ltd (2014). Britest Limited. Introductory training material provided to Newcastle University. England.

Bruscoa, M.J, Steinley, D, Dennis, J (2012) *Emergent clustering methods for empirical OM research*. Journal of Operations Management 30. 454–466.

Burton, H. (1968). *Deposits from whole milk in heat treatment plant- a review and discussion*. Journal of dairy research, 35, 317-330.

Bustamante, P, Pena, M.A and Barra, J. (2000) International Journal of Pharm 194 117-124

Carr, W. A (2011a) Britest cleaning survey II. Engineering Doctorate First year report.

Carr, W. A (2011b) Poster presented at Britest members day, Reebok Stadium, Bolton. UK.

Catchpole, P (2009) Database Navigation Page TSB Zeal Benchmark Tool User Guide Ecolab Ltd.

Changani, S.D, Belmar-Beiny, M.T and Fryer, P.J (1997). *Engineering and Chemical Factors Associated with Fouling and Cleaning in Milk Processing*. Experimental Thermal and Fluid Science. 14. 392 – 406

Chen, M.J, Zhang, Z and Bott, T.R. (1998) *Direct measurement of the adhesive strength of biofilms in pipes by micromanipulation*. Biotechnology Techniques, 12, 875-880.

Clark, J. (2004) *Introducing amines article*. (Accessed online January 2016). Available at http://www.chemguide.co.uk/organicprops/amines/background.html (January 2016).

Clayden J, Greeves, N, Warren, S and Wothers, P (2001) Organic Chemistry. Oxford University Press Inc. New York.

Clint, J. H (1992) *Surfactant Aggregation*. Springer Science and Business Media. New York. (Accessed January 2016). Available online at https://books.google.co.uk/books?hl=en&lr=&id=SczoCAAAQBAJ&oi=fnd&pg=PA1&dq=Surface+Tension+is+the+tension+of+the+surface+film+of+a+liquid+which+is+caused+by+the+attraction+of+the+particles+in+the+surface+layer+by+the+majority+of+the+liquid.+This+tends+to+minimise+the+surface+area&ots=-tiqEYkHne&sig=RJ1rwU97j9Qa5t6BCIpw9LAAWM0#v=onepage&q&f=false

CMR, 2013. *CMR International pharmaceutical R&D fact book*. Thomas Reuters™ [Online] (Accessed 21 07 2014). Available at http://cmr.thomsonreuters.com/pdf/fb-exec-2013.pdf

Cole, P.A, Asteriadou, K Robbins, P.T, Owen, E.G, Montague, G.A and Fryer, P.J. (2010) *Comparison of cleaning of toothpaste from surfaces and pilot scale pipework*. Food and Bioproducts Processing 88. 392 – 400.

Comanor , W.S. and Scherer, F.M., 2012. *Mergers and innovation in the pharmaceutical industry*. Journal of Health Economics, 32. Pp106-113.

Company 1 (2011a) Information provided at site meeting Sept 2011.

Company 1 (2011b) Information provided at a site meeting Sept 2011.

Company 3 (2012) information provided during site visits.

Constantinou, L and Gani, R (1994) *New Group Contribution Method for Estimating Properties of Pure Compounds*. AIChE Journal October 1994 Vol. 40, No. 10 Pg1697 -1710.

Correa, A, Comesana, J.F and Sereno, A.M (1994) *Use of analytical solutions of groups (ASOG) contribution method to predict water activity in solutions of sugars, polyols and urea*. International Journal of Food Science & Technology. Volume 29, Issue 3, pages 331–338.

Cramer, J. (1972) *Evaluation methods for soil removal and soil redeposition*. In Cutler W.G (Ed) Detergency Theory and Test Methods. Part 1. Marcel Dekker, New York, p324-411.

Cumming, J, Hall, C, Harwood, C, Gammage, K (2002) *Motivational orientations and imagery use: a goal profiling analysis*. Journal of Sports Sciences, 01 January 2002, Volume 20(2), pages 127-136.

Daemmrich, A. and Mohanty, A., 2014. *Healthcare reform in the United States and China: pharmaceutical market implications*. Journal of Pharmaceutical Policy and Practice, 7:9.

Décréau, P.M.E. ; Le Guirriec, E. Rauch, J.L. Trotignon, J.G. Canu, P. Darrouzet, F. Lemaire, J. Masson, A. Sedgemore, F. André, M. (2005) *Density irregularities in the plasmasphere boundary layer: Cluster observations in the dusk sector*. Advances in Space Research, Vol.36 (10), pp.1964-1969.

Della Porta, G, Volpe, M.C and Reverchon, E. (2006) *Supercritical cleaning of rollers for printing and packaging industry*. Journal of Supercritical Fluids 37 409-416.

Dong Dong and McAvoy, T.J (1996) *Batch Tracking via Nonlinear Principal Component Analysis*. AIChE Journal. Volume 42, Issue 8 Pages 2199–2208.

Dong, Dong-L, Wu, Qiang ; Zhang, Rui ; Song, Ying-Xia ; Chen, Shu-Ke ; LI, Pei ; Liu, Shou-Qiang ; Bi, Cen-Cen ; LV, Zhen-Qi ; Huang, Song-Lin (2007) *Environmental Characteristics of Groundwater: an Application of PCA to Water Chemistry Analysis*. Yulin Journal of China University of Mining and Technology. Vol. 17(1), page 73-77.

Dotterer, S, Forbes, R.A and Hammill, C (2011) *Impact of metal-induced degradation on the determination of pharmaceutical compound purity and a strategy for mitigation*. Journal of Pharmaceutical and Biomedical Analysis 54 (2011) 987-994.

D. T. Stanton, T.W. Morris, S. Roychoudhury and C.N Parker (1999) *Application of Nearest Neighbour and Cluster Analyses in Pharmaceutical Lead Discovery*. J. Chem. Inf. Comput. Sci39 (1), pp 21-27

Durkee, J (2004a) *Solubility Parameters Part 1*. Cleaning Times. Metal Finishing. (Accessed online June 2014). Available at http://www.metalfinishing.com.

Durkee, J (2004b) *Solubility Parameters Part 2*. Cleaning Times. Metal Finishing. Accessed online June 2014) Available at http://www.metalfinishing.com.

Durkee, J. (2006). *Management of Industrial Cleaning Technology and Processes*. Elsevier Oxford, UK.

Durkee, J.B II (2014) *Cleaning with Solvents: Science and Technology*. Elsevier, Amsterdam.

Dürr, H and Graßhff, A (1999). *Milk Heat Exchanger Cleaning: Modelling of Deposit removal*. Institution of chemical engineers Trans IChemE, Vol 77, Part C.

D. Xu, N. Redman-Furey (2007) *Statistical cluster analysis of pharmaceutical solvents*. International journal of Pharmaceutics, Vol 339, issues 1-2, pages 175-188

EFPIA, 2014. *The pharmaceutical industry in figures: key data 2014*. (Accessed 21 07 2014.Online). Available at: http://www.efpia.eu/uploads/Modules/Mediaroom/figures-2014-final.pdf Envirowise (2008) *Cost Effective Vessel Washing – The Good Practise Guide*.

FDA, (1999) Guidance for Industry. Abbreviated New Drug Applications (ANDA's*). Impurities in Drug Substances* U.S. Department of Health and Human Services Food and Drug Administration Centre for Drug Evaluation and Research (CDER) *J:\!GlJlDANC\2452FNL. WPD*

Fox, Marye, and Whitesell, James K., (2016) *Organic Chemistry* 3rd Edition Interactive Glossary. (Accessed online January 2016) Available at: http://physicalscience.jbpub.com/orgo/interactive_glossary_showterm.cfm?term=Polarizability

Fredenslund A., Jones R.L and Prausnitz J.M.(1975) *Group-Contribution Estimation of Activity Coefficients in Non ideal Liquid Mixtures*, AIChE J., 21(6), 1086-1099

Fryer, P.J and Asteriadou, K (2009). A *Prototype cleaning map: A classification of industrial cleaning processes*. Trends in Food Science and Technology. 20. 255-262.

Gharagheizi, F, Torabi, A.M and Macromol, J. (2006) Sci B Phys 45 285-290.

Gharagheizi, F, Eslamimanesh, A, Mohammadi, A.H and Richon, D (2011) *Group contribution-based method for determination of solubility parameter of non-electrolyte organic compounds*. Ind. Eng. Chem. Res. 2011. Vol.50 p10344 – 10349.

Gillham, C.R, Fryer, P.J, Hasting, A.P.M and Wilson, D.I. (1999). *Cleaning in place of whey protein fouling deposits: Mechanisms controlling cleaning*. Institution of Chemical Engineers. Trans iChemE Vol 77 Part C June 1999.

Gmehling, J, Fischer, K, Li, J and Schiller, M (1993) *Status and results of group contribution methods*. Pure and Applied Chem. Vol. 65, No. 5 p919 – 926.

Görög, S, Babják, M et al (1997) *Drug Impurity profiling strategies*. Talanta 44 (1997) 1517-1526.

Görög, S (2005) *The Sacred Cow: the questionable role of assay methods in characterising the quality of bulk pharmaceuticals*. Journal of Pharmaceutical and Biomedical Analysis 36 (2005) 931-937.

Goverdhan, G, Reddy, A.M et al (2009*) Identification, characterisation and synthesis of impurities of zafirlukast*. Journal of Pharmaceutical and Biomedical Analysis. 49 (2009) 895-900.

Grasshoff, A. (1999). *Fundamental trials to clean UHT tubular modules. Fouling and cleaning in food processing '98, EUR18804 p238-245*. Luxembourg Office for European communities.

Gordon A.D (1995) *A Q survey of constrained classification*. Computational Statistics and Data Analysis 21 17-29.

Goverdhan, G, Reddy, A.M et al (2009*) Identification, characterisation and synthesis of impurities of zafirlukast*. Journal of Pharmaceutical and Biomedical Analysis. 49 (2009) 895-900.

Guidi, L, Ibanez, F, Calcagno, V and Beaugrand, G (2009). *A new procedure to optimize the selection of groups in a classification tree*. Applications for ecological data Ecol. Model., 220 page 451–461.

Hanley, J, A (1983) *Appropriate uses of Multivariate analysis*. Annu. Rev. Public Health 4.155-180.

Hansen, C.M (2007). *Hansen Solubility Parameters: A users handbook*. 2nd ed., CRC Press, Boca Raton.F.L.

Harrington, J. (2001). *Industrial cleaning technology*. Kluwer Academic Publishers, The Netherlands.

Hattori, M. Ikebe, Y. Asaoka, I. Takeshima, T. Bohringer, H. Mihara, T. Neumann, D. M. Schindler, S. Tsuru, T. Tamura, T (1997) *A dark cluster of galaxies at Redshift z=1 (galaxy cluster observation)*. Nature, July 10, Vol.388 (6638) page 146(3).

Hildebrand, J.H and Scott, R.L (1949). *The Solubility of Non-Electrolytes*. 3rd ed. Dover, New York.

Hodgett, R.E., 2013. *Multi-Criteria Decision-Making in Whole Process Design*. PhD Thesis, Newcastle University, England.

Holderbaum T., Gmehling J (1991) *PSRK: A Group-Contribution Equation of State based on UNIFAC"*, Fluid Phase Equilib., 70, 251-265.

Hotelling, H. (1933). *Analysis of a complex of statistical variables into principal components*. Journal of Educational Psychology, 24:417–441.

Housecroft, C.E and Sharpe, A.G (2005) *Inorganic chemistry* (2nd Ed). Pearson. Prentice Hall. England.

Hubert, L (1973*) Monotone invariant clustering procedures*. Psychometrika 38(1): 47-62.

ICH Q2 (1994) *Validation of Analytical Procedures: Text and Methodology*. Current Step 4 version. International conference on harmonisation of technical requirements for registration of pharmaceuticals for human use. ICH Harmonised Tripartite Guideline. (Accessed online 2015) Available at ICH website www.ich.org/

ICH Q3A (2006) *Impurities in New Drug Substances. Current Step 4 version*. International conference on harmonisation of technical requirements for registration of pharmaceuticals for

human use. ICH harmonised Tripartite Guideline. (Accessed online 2015) Available at ICH website www.ich.org/

ICH Q4B Annex 3 (R1) (2010) *Test for particulate contamination: sub-visible particles general chapter*. Current Step 4 version. International conference on harmonisation of technical requirements for registration of pharmaceuticals for human use. ICH harmonised Tripartite Guideline. (Accessed online 2015). Available at ICH website www.ich.org/

ICH Q6A (1999) *Specifications: Test procedures and acceptance criteria for new drug substances and new drug products: Chemical substances*. Current Step 4 version. International conference on harmonisation of technical requirements for registration of pharmaceuticals for human use. ICH harmonised Tripartite Guideline. (Accessed online 2015). Available at ICH website www.ich.org/

ICH Q7 (2000) *Good Manufacturing Practice guidelines for active pharmaceutical ingredients*. Current Step 4 version. International conference on harmonisation of technical requirements for registration of pharmaceuticals for human use. ICH harmonised Tripartite Guideline. (Accessed online 2015). Available at ICH website www.ich.org/

ICH Q8 (2009) *Pharmaceutical development*. Current Step 4 version. International conference on harmonisation of technical requirements for registration of pharmaceuticals for human use. ICH harmonised Tripartite Guideline. (Accessed online 2015). Available at ICH website www.ich.org/

ICH Q10 (2008) *Pharmaceutical Quality System*. Current Step 4 version. International conference on harmonisation of technical requirements for registration of pharmaceuticals for human use. ICH Harmonised Tripartite Guideline. (Accessed online 2015.) Available at ICH website www.ich.org/

ICH Q11 (2012) *Development and Manufacture of drug substances (Chemical entities and Biotechnological/Biological entities).* Current Step 4 version. International conference on harmonisation of technical requirements for registration of pharmaceuticals for human use. ICH Harmonised Tripartite Guideline. (Accessed online 2015). Available at ICH website www.ich.org/

Irigoin, J, Paladino, I, Civeira, G, Costa, M.C (2016) *Physical and chemical variables analysis for clustering of soils in the longitudinal dunes of Sandy Pampa, Argentina*. Environ Earth Sci. 75: 1196.

Ishiyama, E.M, Henis, A.V, Paterson, W.R, Spinelli, L and Wilson, D.I. (2010). *Scheduling cleaning in a crude oil preheat train subject to fouling; Incorporating desalter control*. Applied Thermal Engineering 30 1852 – 1862.

Járvás, G, Quellet, C and Dallos, A.(2011) *Estimation of Hansen solubility parameters using multivariate nonlinear QSPR modelling with COSMO screening charge density moments*. Fluid Phase Equilibria 309 (2011) 8-14.

Izenmann, A. J (2013) *Modern Multivariate Statistical techniques*. Regression, Classification, and Manifold learning. (2nd Ed) Springer. New York.

Jensen, B. B. B, Stenby, M and Nielsen, D, F (2006) *Improving the cleaning effect by changing average velocity*. Trends in Food Science and Technology 18 (2007) S52-S63

Jolliffe, I. T (2002), *Principal Component Analysis (2nd Edn.)*. Springer- Verlag New York, Inc United States of America.

Joyce, M. Chaboyer, B (2015) *Investigating the consistency of stellar evolution models with globular cluster observations via the red giant branch bump*. The Astrophysical Journal, Vol.814(2), page 142 (11pp).

Jung, Y, Park, H, Drake, B.L and Du, D (2002) *A decision criterion for the Optimal Number of Cluster in Hierarchical Clustering*. Kluwer Academic Publishers.

Juran, J.M., 1992. *Juran on quality by design: the new steps for planning quality into goods and services*. New York: The Free Press.

Kaiser, H.F (1960). *The application of electronic computers to factor analysis*. Educational and Psychological Measurement, 20, 141-151.

Khanna, I., 2012. *Drug discovery in pharmaceutical industry: productivity challenges and trends*. Drug Discovery Today, 17 (19/20).

Krishna Reddy, K.V.S.R, Moses Babu, J et al (2002) *Isolation and characterisation of process-related impurities in rofecoxib*. Journal of Pharmaceutical and Biomedical Analysis 29 (2002) 355- 360.

Kühne, R, Ubert, R-U, Kleint, F, Schmidt, G and Schüürmann, G (1995). *Group Contribution methods to estimate water solubility of organic chemicals*. Chemosphere. 30. 22. 2061-2077.

Lainez, J.M., Schaefer, E. and Reklatis, G.V., 2012. *Challenges and opportunities in enterprise-wide optimization in the pharmaceutical industry*. Computers and Chemical Engineering, 47, pp19-28.

Leclercq-Perlat, M.N, Tisser, J.P and Benezech, T. (1993) *Cleanability of stainless steel in relation to chemical modifications due to industrial cleaning procedures used in the dairy industry*. Journal of Food Engineering. 23. 449 – 465.

Leite, A, Souza Figueiredo, P, Caracas, H, Sindeaux, R, Guimarães, A, Lazarte, L, Paula, A, Melo, N (2015) *Systematic review with hierarchical clustering analysis for the fractal dimension in assessment of skeletal bone mineral density using dental radiographs*. Oral Radiology, Vol.31 (1), page 1-13.

Liu, W. Christian, G.K Zhang, Z and Fryer, P.J (2006a). *Direct measurement of the force required to disrupt and remove fouling deposits of whey protein concentrate*. International Dairy Journal. 16. 164-172.

Liu, W, Fryer, P.J, Zhang, Z, Zhao, Q and Liu, Y. (2006b). *Identification of cohesive and adhesive effects in the cleaning of food fouling deposits*. Innovative Food Science and Emerging Technologies. 7. 263 – 269.

Liu, W, Zhang, Z and Fryer, P.J (2006c). *Identification and modelling of different removal modes in the cleaning of a model food deposit*. Chemical Engineering Science 61 7528 – 7534.

Määttä, J and Kymäläinen (2011) *Application of radiochemical determination methods in cleanability research of building materials*. Journal of Environmental Radioactivity 102 (2011) 649-658.

Macnaughton-Smith P, Williams W.T, Dale M.B. and Mockett L.G (1964) *Dissimilarity analysis: a new technique of hierarchical sub-division*. Nature 202: page 1034-1035.

Maere, T, Villez, C, Marsili-Libelli, S, Naessens, W, Nopens, I (2012) *Membrane bioreactor fouling behaviour assessment throughprincipal component analysis and fuzzy clustering*. Water research 46. 6132e6142.

Malinowski, E.R., Weiner, P.H., Levinstone, A. R. (1970) *Factor analysis of solvent shifts in proton magnetic resonance* J. Phys. Chem., 1970, *74* (26), pp 4537–4542.

Massey, A. G. (1990) *Main Group Chemistry*. Ellis Horwood. New York.

Melnikov, N. N (1971) *Chemistry of pesticides*. Editor Gunther, F.A and Gunther. J.D Springer-Verlag. New York 1st Edition. (Accessed online January 2016) Available at https://books.google.co.uk/books?id=OVjhBwAAQBAJ&pg=PA401&lpg=PA401&dq=o-heterocyclic+compounds+warfarin&source=bl&ots=YTtKQzQNOm&sig=hrKd1Tsapc3uG03lGqNzCnC0Kl8&hl=en&sa=X&ved=0ahUKEwiU4rX-3-PLAhVD8RQKHUVGCbE4ChDoAQgdMAE#v=onepage&q=o-heterocyclic%20compounds%20warfarin&f=false

Miller, J.G, Roth, A.V, 1994. *A taxonomy of manufacturing strategies*. Management Science 40 (3), 285–304.

Minitab (2016) StatGuide Definition of Principal Components Minitab version 16. N.A. Minitab Inc.

MSD (2015). *Macrolides*. Merck Vet Manual. Merck Sharpe and Dohme Corp. (Accessed online January 2016). Available at http://www.merckvetmanual.com/mvm/pharmacology/antibacterial_agents/macrolides.html

NHS (2013) *National Health Service England and Wales Drug Tariff*. The Stationary Office London.

Nicholls. A. O, McIntyre, S and Stol, J (2010) *Comments on optimising the selection of the number of groups in a classification tree*. Ecological Modelling Vol. 221. Issue 9 page 1333-1335.

Nomikos, P., and J. MacGregor, (1994) *Monitoring of Batch Processes Using Multi-way PCA*, AIChE J., 40, 1361.

Nur, T, Azira Y.B, Che Man R. N et al (2014) *Use of principal component analysis for differentiation of gelatine sources based on polypeptide molecular weights*. Food Chemistry. Volume 151, Pages 286–292.

Oliveira, A.C., Dos Santos, V.S., Dos Santos, D.C., Carvalho, R.D.S, Souza, A.S., Ferreira, S.L.C., (2014) *Determination of the mineral composition of Caigua (Cyclanthera pedata) and evaluation using multivariate analysis*. Food Chemistry, 1 June 2014, Vol.152, page 619-623.

O'Neil, M. J. (2013) (ed**.).** *The Merck Index - An Encyclopedia of Chemicals, Drugs, and Biologicals*. Cambridge, UK: Royal Society of Chemistry, 2013, p. 796.

Paul, S.M., Mytelka, D.S., Dunwiddie, C.T., Persinger, C.C., Munos, B.H., Lindborg, S.R. and Schacht, A.L., 2010. *How to improve R&D productivity: the pharmaceutical industry's grand challenge*. Nature Reviews, Drug Discovery, 9. Pp203-214.

Pelczaraska, A, Ramjugernath, D, Rarey, J, and Domańska, U (2013). *Prediction of the solubility of selected pharmaceuticals in water and alcohols with a group contribution method*. Journal of Chemical Thermodynamics Vol 62, 118-129

Peng, X, Li, X, Shi, X and Guocor, S (2014) *Evaluation of the aroma quality of Chinese traditional soy paste during storage based on principal component analysis*. Food Chemistry. Volume 151, Pages 532–538.

Pesonen-Leinonen, E and Redsven, I et al (2006) *Determination of soil adhesion to plastic sufaces using a radioactive tracer*. Applied Radiation and Isotopes 64 (2006) 163-169.

Prasanna, S and Doerksen R.J (2009) *Topological polar surface area: a useful descriptor in 2D-QSAR*. Curr Med Chem 16 (1) 21-41. (Accessed online January 2016). Available at http://www.ncbi.nlm.nih.gov/pubmed/19149561 in .

Preveena, S.M, Kwan, O.W and Aris, Z. A (2012) *Effect of data pre-treatment procedures on principal component analysis: a case for mangrove surface sediment datasets*. Environ Monit. Asses. 184:6855 6868.

Pritchard, N.J. de Groerden, G and Hastings, A.P.M. (1988). *The removal of milk deposits from heated surfaces by improved cleaning processes*. Proc Fouling in Process Plant., St. Catherine's college, Oxford,465-479.

Prosek, M, Krizman, M and Kovac, M. (2005). *Evaluation of a rinsing-based cleaning process for pipes*. Journal of Pharmaceutical and Biomedical Analysis. 38. 205 – 513.

Pubchem - Open chemistry database, Compound Summary for CID 2244. National Centre for Biotechnology Information. (Accessed January 2016). Available at https://pubchem.ncbi.nlm.nih.gov/compound/aspirin.

Qin, C and Granger, A et al (2009) *Quantitative determination of residual active pharmaceutical ingredients and intermediates on equipment surfaces by ion mobility spectrometry*. Journal of Pharmaceutical and Biomedical Analysis 51 (2010) 107-113

RB Plant 2016 (Accessed online January 2016) Website http://www.rbplant.com/

Reach-serve (2016) *Regulation, Registration, Evaluation, Authorisation and Restriction of Chemicals*. Physical chemical property definitions. (Accessed online January 2016). Available at http://www.reach-serv.com/index.php?option=com_content&task=view&id=59&Itemid=129

Reid, M.K and Spencer K.L (2009*) Use of principal components analysis (PCA) on estuarine sediment datasets: The effect of data pre-treatment*. Environmental Pollution Volume 157, Issues 8–9, pages 2275–2281.

Rencher, A.C (2002) *Methods of Multivariate Analysis (2nd Ed.).* John Wiley and Sons, Inc. Canada.

Resto, W and Darimar et al (2007) *Cleaning Validation 2: Development and validation of an ion chromatographic method for the detection of traces of CIP-100 detergent*. Journal of Pharmaceutical and Biomedical Analysis 44 (2007) 265-269.

Roberts, R.J and Rowe, R.C. (1993) International Journal of Pharm 99 157-164

Roessling, G (2011). *Points to consider for Biotechnology Cleaning*. Validation Parenteral Drug Association Training Course. Bordeaux, France.

Roux, M (1991), Devillers J. and Karcher W. (Eds). *Basic procedures in hierarchical cluster analysis.* Applied Multivariate Analysis in SA–R and Environmental Studies. Kluwer Academic Publishers. Dordrecht. Page 115-135.

Roy, J (2002) *Pharmaceutical Impurities- A Mini Review*. AAPS PharmSciTech 2002; 3 (2) article 6. Accessed online January 2016. Available at www.aapspharmscitech.org.

RSC. 2014. Royal society of Chemistry. *Science needs more funding to keep Britain competitive*. Press Release. Accessed online 21 July 2014. Available at http://www.rsc.org/AboutUs/News/PressReleases/2013/Chemistry-We-Mean-Business-report.asp.

Savova, M, Kolusheva, T, Stourza, A and Seikova, I (2007). *The use of group contribution method for predicting the solubility of seed polyphenols of vitis viniferal within a wide polarity range in solvent mixtures*. Journal of the University of Chemical Technology and Metallurgy. 42, 3, p295 – 300.

Schäfer. S et al (2015) *Bioaccumulation in aquatic systems: methodological approaches, monitoring and assessment*. Environmental Sciences Europe Bridging Science and Regulation

at the Regional and European Level 2015 **27**:5 (Accessed on line January 2016 open access article) Springer. Available at http://enveurope.springeropen.com/articles/10.1186/s12302-014-0036-z

Schlecht, H.P and Bruno, C (2016) *Macrolides*. MSD Manual MSD Merk Sharpe and Dohme Corp. Merck and Co., Inc., Kenilworth, NJ, USA. (Accessed online January 2016). Available at http://www.msdmanuals.com/en-gb/professional/infectious-diseases/bacteria-and-antibacterial-drugs/macrolides

Shasun (2012) information provided during site visits.

Shebs, W. (1987) *Radioisotope techniques in detergency*. In Gutler, W.G (Ed) Detergency Theory and Technology. Surfactant Science Series, vol. 20. Dekker, New York, p126-187.

Simmons, M.J.H, Jayaraman, P and Fryer, P.J. (2007). *The effect of temperature and shear upon the rate of aggregation of whey protein and its implications for milk fouling*. Journal of Food Engineering 79, 517-528

Singh, K. P. Malik, A, Mohan, D, Sinha, S (2004) *Multivariate statistical techniques for the evaluation of spatial and temporal variations in water quality of Gomti River India - a case study*. Water Research, November. Vol.38 (18), page 3980-3992.

Sneath P.H.A. and Sokal R.R (1973). *Numerical taxonomy*. W.H. Freeman and Co. San Francisco. Page 573.

Sørlie, T, Tibshirani, R, Parker, J, Hastie§, T, Marron, J. S, Nobel, A, Deng‖, S, Johnsen, H, Pesich, R, Geisler, S, Demeter, J, Perou,C.M, Lønning P.E, Børresen-Dale, A, and Botstein, D (2003) *Repeated observation of breast tumor subtypes in independent gene expression data sets*. Proceedings of the National Academy of Sciences of the United States of America Issue Vol. 100. no.14.

Steris (2012) Information on products. Accessed online 2014) Available at http://www.steris.com/

Talford, M. (2009) Plant Cleaning Survey v1. Britest Ltd.

Usanova, M.E, Darrouzet, F, Mann, I.R and Bortnik, J (2013) *Statistical analysis of EMIC waves in plasmaspheric plumes from Cluster Observations*. Journal of Geophysical Research – Space Physics Volume 118, Issue 8 pages 4946-4951.

Van Asselt, A.J. Van Houwelingen, G and Giffel, M.C. TE. (2002). *Monitoring system for improving cleaning efficiency of cleaning-in-place processes in dairy environments*. Institution of Chemical Engineers. Trans IChemE, Vol 80, Part C. Dec.

Van Roosmalen, M.J.E, Woerlee, G.F and Witkamp, G.J (2004) *Surfactants for particulate soil removal in dry- cleaning with high pressure carbon dioxide*. Journal of supercritical fluids 30 97 – 109.

Vialle, C, Sablayrolles, C, Lovera, M, Jacob, S, Huau, M.-C. Montrejaud-Vignoles, M (2011) *Monitoring of water quality from roof run off: Interpretation using multivariate analysis*. Water Research, Volume 45(12), page 3765-3775.

Villegas, R, Salim, A, Collins, M, Flynn, A, Perry, I, Perry (2004) *Dietary patterns in middle aged Irish men and women defined by cluster analysis*. Public Health Nutrition, Volume 7(8), pages 1017-1024.

Visser, J. (1995). *Adhesion and removal of particles I and II*. In Melo, L. F., Bottt, T. R., and Bermardo (Eds). Fouling Science and technology. Dordrecht: Kluwer Academic Publishers.

Wallace, J. Champagne, P. Hall, G. (2016) *Multivariate statistical analysis of water chemistry conditions in three wastewater stabilization ponds with algae blooms and pH fluctuations*. Water Research, Vol. 96, page 155-165.

Ward J.H. Jr (1963) *Hierarchical grouping to optimize an objective function*. J. Am. Stat. Assoc. 58: 236-244.

Wathan, S (1995) *Manufacturing strategy in business units: an analysis of production process focus and performance*. International Journal of Operations and Production Management 15 (8), 4–13.

Webster, R (2001) *Statistics to support soil research and their presentation*. European journal of soil science Volume 52 Issue 2 Pages 331-340.

Wold, S (1987) *Principal Component Analysis*. Chemometric and Intelligent Laboratory Systems 2 37-52 Elsvier Science Publishers.

Yu, L.X., 2008. *Pharmaceutical quality by design: product and process development, understanding, and control*. Pharm Res, 25, pp.781-791.

Zayas, J and Colón, H et al (2006) *Cleaning Validation of a chromatographic method for the detection of traces of LpHse detergent*. Journal of Pharmaceutical and Biomedical Analysis 41 (2006) 589-593.

Zitko,V (1994) *Principal Component Analysis in the Evaluation of Environmental Data*. Marine Pollution Bulletin Vol. 28 No12 page 718-722.

Computer Software

Chemspider. Royal Society of Chemistry [Computer Software]. (Accessed online 2013, 2014 and 2015). Available at www.chemspider.com

ChemDraw (version 14) [Computer Software]. Accessed online 2013 and 2014. Perkin Elmer.

Excel (version 2007) [Computer Software]. Microsoft Corporation.

Minitab (version 16). (2015) [Computer Software].N.A. Minitab Inc.

Appendix I:

Plant Cleaning Survey

Introduction

The purpose of this survey is to understand the current approaches to plant cleaning taken by Britest member companies in order to provide common background for further work on the plant cleaning project, building on the survey of Britest Members conducted in 2009. You may wish to print out the survey and complete by hand, or you may edit this document to include your responses. Any information provided will be treated as confidential between the Member, Britest, and the Newcastle University Plant Cleaning EngD (Wendy Carr); All results will be made anonymous before any analysis circulated within the Britest membership.

Questions from the original survey are in black text. Additional questions to help provide further information are in blue text.

Please provide the following details – we may wish to follow up specific points with you.

Organisation:

Contact details:

Nature of process plants (mark all that apply):

Multipurpose batch Y / N Single product batch Y / N Continuous Y / N

Can you indicate which product type you are involved in manufacturing?

Chemical

Pharmaceutical

Biopharmaceutical

Other (please specify):

Plant cleaning protocols

1. With an emphasis on one process please provide a brief description of how your current plant cleaning protocols are/were developed, including an indication of where in the process lifecycle they were considered and an outline of any tools/methods used in their development.

2. On what priority is the cleaning protocol based?

   The type of contaminants Y / N

   The type of product Y / N

The type of plant equipment Y / N

The level of soiling Y / N

Pharmaceutical or industry standards or requirements Y / N

A combination of the above factors with no definite priority Y / N

Other (please specify):

3. Are your cleaning protocols based on utilising:

   Volume of cleaning agent (e.g. solvent, detergent, water) used? Y / N

   Contact time for the cleaning agent? Y / N

   Removal of contaminant(s) to specified levels? Y / N

   Other (please specify):

4. How many contaminants are you typically trying to remove?

   1-4 5-9 10 or more

5. Are your cleaning protocols developed from an understanding of the contaminant types? Y / N

   If yes please specify and give examples:

6. Which is the priority, level of contamination or type of contamination?

   Level

   Type

   Both equally important

7. What are the main contaminant types in your typical processes?

   Chemical based

   Biological based

   Residual cleaning agents

   Other (please specify):

8. Do you typically clean at:

   Ambient temperature?Y / N

   Elevated temperature?Y / N

9. How do you determine the cleaning temperature?

Suppliers' recommendations Y / N

Lab assessment during development Y / N

In-plant assessment during commissioning Y / N

Other (please specify):

10. Do you clean between batches of the same product? Y / N

11. Do the protocols for cleaning between batches differ from cleanouts between products? Y / N

    If so, please briefly outline the key differences:

Process-specific cleaning issues

Consider a process that you operate which exhibits some cleaning difficulties:

12. Does an increase in the product batch size affect the amount of soiling? Y / N

13. Is there is an increase in the level of soiling between different batches Y / N

14. Is it believed that the soiling is generally variable between batches? Y /N

    If yes, do you think the soiling variability is due to:

    Raw material batch variations

    The types of raw materials used

    Other (please specify):

15. Do you know whether the main plant soiling occurs at one particular stage in the process? Y / N

16. Is the soiling specific to one area of the plant? Y /N

    If yes, please indicate where:

17. Is this area targeted for specific cleaning? Y / N

    If yes, please specify all of which apply:

    Is it taken apart for cleaning?

    Flushed more than the other areas of the plant

Targeted with a specific cleaning agent

Isolated from the other areas of the plant for cleaning

Other (please specify):

18. Have you identified any chemical or biological structures, or physical properties in the contaminants, that have been targeted by the inclusion of a specific cleaning agent in the design of the cleaning protocol? Y / N

    If Yes, please specify the nature of the structure/property and the cleaning agent selected:

19. Has the cleaning protocol been developed to specifically remove any of these contaminants by the inclusion of specific cleaning agents such as acid or alkali?

    Y / N

    If Yes please specify:

20. Can you specify the cleaning agent?

    Is the same cleaning method carried out on other similar processes at your site that manufacture different products Y / N

    If Yes please specify what the process is and if the cleaning is effective:

21. In your opinion can you indicate how effective the current cleaning method is?

    It is very effective

    It is sometimes effective

    It is not very effective

    It is never effective

22. If money, time and other resources were not limited how would you improve or change the cleaning method for this product?

    Briefly explain below:

Plant cleaning methods

23. What methods do you use to clean your process plant?

Specific Clean-in-Place technology	Y / N
Washouts without disassembly of equipment/pipework	Y / N
Manual cleaning with equipment/pipework disassembled	Y / N
Other (please specify):	

24. Do you use any specific cleaning equipment for this process? Y / N

    If yes please specify:

25. Has any equipment been specifically designed for cleaning post manufacturing processes Y / N

    If yes please specify:

Cleaning agents

26. What cleaning agents do you use?

Organic Solvents	Y / N
Aqueous detergent	Y / N
Mineral acid/alkali	Y / N
Water	Y / N

27. How are your cleaning agents selected?

Suppliers' recommendation	Y / N
Lab assessment during development	Y / N
In-plant assessment during commissioning	Y / N
Other (please specify)	

28. Do you use combinations of cleaning agents for individual processes (e.g. solvent cleaning followed by water washes)? Y / N

    Please provide a brief outline:

29. Do you recover cleaning agents used in cleaning operations:

For process use? Y / N

For use in future cleaning operations? Y / N

30. Are cleaning agents are recovered post cleaning processes? Y / N

If yes please indicate what is recovered:

31. Is there an identifiable reason that solvent recovery is not carried out. Is this due to any factors indicted below? Please indicate all that apply.

Lack of solvent storage

Level of soil in the solvent

Type of soil in the solvent

Never considered it

Other (please specify):

32. Do you recover the contaminants removed for further processing? Y / N

Analysis and validation

33. Please briefly outline how you validate your cleaning protocols:

34. How do you validate plant cleanliness post-cleaning?

Analysis of surface swabs Y / N

Analysis of cleaning agent effluent Y / N

Analysis of rinse effluent Y / N

Visual inspection Y / N

Other (please specify):

35. Please briefly outline the analytical techniques you use to validate cleanliness:

36. What is the reasoning for using your current validation technique? Please indicate all that apply.

    Based on a standard industry technique for this process?

    Is it driven by a pharmaceutical standard?

    Is it based upon an FDA requirement?

    It is all that is available at this site

    Other (please specify):

37. Do you think that this technique is an effective method for validating cleanliness?

    Y / N

38. Are there specific reasons that you do not use an alternative validation technique?

    Please indicate all that apply:

    We are not aware there is another method for this process

    Introducing a new method is not cost effective

    Current method is the best for our requirements

    No time to validate a new method

    Regulatory restrictions

    Company protocol

    Other please specify:

39. Please briefly outline how you determine your acceptance criteria:

40. What course of action do you typically take if analysis indicates the plant is not cleaned to the acceptance criteria?

    Full plant cleaning to the standard protocol Y / N

    Targeted cleaning using a subset of the standard protocol Y / N

    Targeted cleaning following a further protocol Y / N

    Other (please specify):

41. Can you give an indication of the effectiveness of the current cleaning protocol? Does this method work:

On the first attempt	Every time / more than 50% of the time / less than 50% of the time
On the second attempt	Every time / more than 50% of the time / less than 50% of the time
More than 2 attempts	Every time / more than 50% of the time / less than 50% of the time

Time and cost

42. For a typical product cycle/campaign please indicate:

    Batch cycle time, days;

    Between batch cleaning time, days;

    Total production time, days;

    Post campaign cleaning time, days;

43. What is the typical ratio of volume of cleaning agent to total volume of plant?

44. Does this depend on the type of plant equipment? Y / N

    If Yes please explain:

    Does this depend on the size of the plant? Y / N

    If Yes please explain:

45. For a typical product what is the approximate cost of the cleaning activity as a percentage of product cost?

    <5% 6-15% 17-25% >25%

46. Please identify which of these the largest contributor to the cost of the cleaning activity:

    Plant downtime Labour Energy Cleaning agent

47. Can you indicate the potential cost of downtime incurred by ineffective cleaning methods as a percentage of the product costs

    <5% 6-15% 17-25% >25%

48. As a percentage of total plant downtime what percent is attributable to cleaning activity

    <5% 6-15% 17-25% >25%

    >50% >70%

49. What would be your preferred approach to reduce process downtime associated with cleaning? Please give an illustrative example:

General Questions

50. Can you give your definition of clean with reference to your plant?

51. In your opinion when would you state that you know something is clean? Please indicate all that apply.

    It looks visibly clean

    Is validated as clean but there is some soiling visible

    Is validated as clean but there are stains visible

    The vessel is only clean when the validation techniques indicate cleanliness

    Other

52. If a vessel or piece of equipment is marked or stained is stain mapping carried out?

    Y / N

    If Yes, please indicate what instrumentation is used to carried this out

53. If the stains are permanent have you tried to remove them? Y / N

    If yes how was this carried out? Please specify:

    Was this effective? Y / N

    Please explain the above answer:

54. Is the stain a result of a known contaminant or a result of the cleaning process?

    Please indicate below

55. Would the use of disposable equipment be effective on the plant to remove the need for cleaning? Y / N

56. If the answer to question 55 is No, is this due to factors indicated below? Please indicate all that apply.

    Disposal technology is not suitable for this process

    Please indicate why this is the case:

    This is not cost effective

    The scale of the process makes the use of disposable technology unfeasible

    Other please specify:

Appendix II:

Active Pharmaceutical Ingredients and Pharmaceutical Products used in this research. Information was obtained from the Pharmaceutical companies' websites in the public domain. The website addresses are also given below.

Shasun

Products manufactured – Brofen (Ibuprofen), Cycloserine, HPMPC (Cidofovir), Isradipine, Ketoprofen, Gabapentin, Quinapril, Ranitidine, Sevelamar.

www.shasun.com/

Abbott

Products manufactured-Blopress (candesartan cilexetil), Calcijiex (Calcitrol), Clarithromycin, Hytrin, Klacid (Clarithromycin),Epival (Sodium valproate), Lupron (Leuproreline), Nimbex (Cisatracurium besilate), Paricalcitol (Zemplar), Progesterone, Severane, Eprosartan (Teveten).

www.abbott.co.uk/

AstraZeneca

Products manufactured – Citanest (Prilocaine), Imdur (Isosorbide mononitrate), Isoflurane, Marcaine (Bupivacaine), Oxis (Formoterol), Plendil (Felodipine).

www.astrazeneca.co.uk/

AMRI (Albany Molecular Reseach Inc.)

Products manufactured – Furosemide, Merperidine, Warfarin.

www.amriglobal.com/

Hovione

Products manufactured – Ciclesonide, Clobetasol propionate, Dexamethosone dipropionate, Halobetasol, Ixodixanol, Iohexol, Iopamidol, Ivermectin, Doxycycline hyclate, Doxycycline monohydrate, Fluticasone furaroate, Fluticasone propionate, Mometasone furoate anhydrous, Mometasone furoate monohydrate, Roxithromycin, Salmeterol xinafoate, Sumatriptan base, Tamsulosin.

www.hovione.com/

Pfizer

Products manufactured – Venlafaxine

www.pfizer.co.uk/

Eli Lilly

Products manufactured – Nizatidine, Olanzapine

www.lilly.co.uk

UCB

Products manufactured – Metolazone

www.ucbpharma.co.uk/

Jhp pharmaceuticals

Products manufactured – Methohexital

www.pharmaceutical-technology.com/

Wyeth

Products manufactured – Meprobamate

www.wyeth.com

Novartris

Products manufactured – Ciclosporin

www.novartris.co.uk

GE

Products manufactured – Gadopentetate dimeglumine, Gadopentetate monomeglumine

www3.gehealthcare.co.uk/

Watson Laboratories

Products manufactured – Folic Acid

Now known as Allergan

www.allergan.com/

Note – Active Pharmaceutical Ingredients (API's) have more than one name listed in some cases. This is due to the fact that some API's have a brand name at some companies and also a generic drug name. Both are listed where appropriate.

Appendix III:

Variables used in Database 2 Physicochemical Analysis

Exact mass, molecular weight, atoms C (Carbon), O (Oxygen), F (Fluorine), H (Hydrogen). Sulphur (S), N (Nitrogen), Cl (Chlorine), Boiling Point both in Kelvin (K) and in °c, Melting point K, Critical Temperature K, Critical Pressure (Bar), Critical Volume (cm3/mol), Gibbs energy (kJ/mol), LogP, MR (cm3/mol) Henry's Law, Heat of Form, tPSA, CLogP, CMR, ACD/LogP, ACD/LogD (ph5.5), ACD/BCF (ph5.5), ACD/KOC (ph5.5), H bond acceptors, Freely rotating bonds, Index of refraction, Molar Volume (cm), Surface Tension dyne/cm, Flash Point, ACD/Log D (ph7.4), ACD/BCF (ph7.4), ACD/KOC (ph7.4), H bond donors, Polar surface area, Molar Refractivity (cm), Polarizability, Density, Enthalpy of vaporisation (kJ/mol), Vapour pressure.

Classes of API including Dermatological, Nasal and inhalation, Injectable, API (because of a lack of other classification given).

Appendix IV:

This appendix shows variables used in database 1 and database 3 which were analysed by PCA.

Functional Groups

Amine functional groups; Primary, Secondary, Tertiary, Aromatic/enamine.

Alcohol OH functional groups; Primary, Secondary, Tertiary, Vinyl alcohol, Phenol .

Acidic functional groups; Carboxylic, Sulfonated, Other

Carbonyl functional groups; Ketone, Aldehyde, Enone, Ester, primary, secondary, tertiary, Anhydride, Epoxide, Thioester

Other Nitrogen groups; Oxime, Oxazolidinone, Urea, Guanidine

Other functional groups; Ether, Sulfonamide, Sulfone, N-Oxide, Nitrile, Thiol, Thioether, Fluorine, Pyridine, Alkyl halide, Aryl halide, Alkene, Nitrate, Nitro, Carbamate, Phosphate, Other, Hydrozone, Phosphonate, Alkylgreater than 5 C

Structural features and Organic framework

Steroid, Hormone, O-heterocyclic, N-heterocyclic, S-heterocyclic, Long alkyl, Phenyl ring, Erythromycin derivative, Tetracycline, Macrocyclic, Macrolide, Benzodiazepine, Barbiturate, Water, Ethanol, HCl, Na+, Gd3+

Appendix V:

Information and results from analysis (Chapter 5)

Principal Component Analysis: Database One: Chemical Functional Group Information.

The Principal Components determined for the variables from database 1 are listed below.

Eigenanalysis of the Covariance Matrix
69 cases used, 2 cases contain missing values

Eigenvalue	4.9682	4.1165	3.6005	3.3085	3.2663	3.0538	2.7193	2.6784
Proportion	0.086	0.071	0.062	0.057	0.057	0.053	0.047	0.046
Cumulative	0.086	0.157	0.220	0.277	0.333	0.386	0.433	0.480

Eigenvalue	2.2458	2.0947	1.8695	1.8321	1.7762	1.5803	1.5134	1.3747
Proportion	0.039	0.036	0.032	0.032	0.031	0.027	0.026	0.024
Cumulative	0.519	0.555	0.587	0.619	0.650	0.677	0.703	0.727

Eigenvalue	1.2593	1.1599	1.0904	1.0540	1.0011	0.9683	0.9409	0.8940
Proportion	0.022	0.020	0.019	0.018	0.017	0.017	0.016	0.015
Cumulative	0.749	0.769	0.788	0.806	0.823	0.840	0.856	0.872

Eigenvalue	0.8392	0.8157	0.7700	0.6100	0.6064	0.5758	0.4579	0.3913
Proportion	0.015	0.014	0.013	0.011	0.010	0.010	0.008	0.007
Cumulative	0.886	0.901	0.914	0.924	0.935	0.945	0.953	0.960

Eigenvalue	0.3387	0.3151	0.2813	0.2317	0.2049	0.1983	0.1556	0.1234
Proportion	0.006	0.005	0.005	0.004	0.004	0.003	0.003	0.002
Cumulative	0.966	0.971	0.976	0.980	0.983	0.987	0.990	0.992

Eigenvalue	0.1052	0.0998	0.0866	0.0498	0.0431	0.0294	0.0272	0.0182
Proportion	0.002	0.002	0.001	0.001	0.001	0.001	0.000	0.000
Cumulative	0.993	0.995	0.997	0.998	0.998	0.999	0.999	1.000

Eigenvalue	0.0112	0.0071	0.0036	0.0000	0.0000	0.0000	0.0000	-0.0000
Proportion	0.000	0.000	0.000	0.000	0.000	0.000	0.000	-0.000
Cumulative	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000

Eigenvalue	-0.0000	-0.0000	-0.0000	-0.0000	-0.0000	-0.0000	-0.0000
Proportion	-0.000	-0.000	-0.000	-0.000	-0.000	-0.000	-0.000
Cumulative	1.000	1.000	1.000	1.000	1.000	1.000	1.000

Eigenvalue	-0.0000
Proportion	-0.000
Cumulative	1.000

Variable	PC1	PC2	PC3	PC4	PC5	PC6	PC7
Primary	0.156	0.186	-0.05	-0.166	-0.122	-0.129	-0.214
Secondary	0.112	0.047	0.169	0.118	0.123	-0.247	-0.022
Tertiary	0.071	-0.069	0.378	0.190	0.108	0.005	0.121
Aromatic/enamine	0.150	0.231	0.041	-0.118	-0.105	0.327	0.011
Primary_1	0.089	0.079	0.080	0.053	0.035	-0.127	0.177
Secondary_1	0.062	-0.021	0.062	0.115	0.108	0.019	-0.097
Tertiary_1	-0.163	-0.231	0.255	-0.244	-0.071	0.003	-0.013
Vinyl alcohol	-0.022	-0.182	0.165	0.137	-0.421	0.013	-0.098
Phenol	0.104	0.137	-0.013	-0.116	-0.063	-0.154	-0.087
Carboxylic	0.095	0.075	0.255	0.206	0.175	-0.202	0.103
Sulfonated	0.000	0.000	0.000	0.000	0.000	0.000	0.000
Other	0.000	0.000	0.000	0.000	0.000	0.000	0.000
Ketone	-0.262	-0.123	-0.202	0.119	-0.030	0.058	-0.020
Aldehyde	-0.000	0.000	-0.000	0.000	-0.000	-0.000	-0.000
Enone	-0.025	-0.009	-0.088	0.044	0.024	-0.033	-0.052
Ester	-0.117	-0.190	-0.221	-0.066	0.081	0.092	0.048
1o amide	0.003	-0.099	0.014	0.052	-0.213	-0.001	0.095
2o amide	0.144	0.181	-0.054	-0.152	-0.147	-0.018	0.212
3o amide	0.065	0.055	-0.067	-0.073	-0.108	-0.035	0.370
Anhydride	0.000	-0.000	0.000	-0.000	0.000	0.000	0.000
Epoxide	0.000	-0.000	0.000	-0.000	0.000	0.000	0.000
Thioester	-0.059	0.009	-0.015	0.139	0.110	0.107	0.082
Oxime	-0.034	-0.200	0.179	-0.313	0.116	0.040	0.022
Oxazolidinone	-0.003	-0.010	-0.026	0.019	0.006	-0.006	0.002
Urea	0.003	0.052	-0.007	-0.015	-0.007	0.018	0.048
Guanidine	0.116	0.143	-0.047	-0.179	-0.164	0.056	0.013
Ether	-0.010	-0.202	0.075	-0.321	0.116	0.040	-0.047
Sulfonamide	0.038	0.020	0.027	0.017	0.036	-0.065	-0.060
Sulfone	0.000	-0.049	-0.031	-0.085	0.044	0.030	0.007
N-Oxide	0.033	-0.009	0.016	0.048	0.092	0.035	-0.311
Nitrile	-0.000	0.000	-0.000	0.000	-0.000	-0.000	-0.000
Thiol	-0.000	0.000	-0.000	0.000	-0.000	-0.000	-0.000
Thioether	0.051	0.134	0.156	0.097	0.070	0.433	0.002
Fluorine	-0.191	-0.021	-0.157	0.106	0.078	0.058	0.037
Pyridine	-0.186	0.126	0.043	-0.049	-0.018	-0.002	0.022
Alkyl halide	-0.110	-0.107	-0.233	0.105	0.034	0.060	0.002
Aryl halide	-0.031	0.141	-0.007	-0.098	-0.044	-0.142	-0.190
Alkene	0.007	-0.013	0.000	-0.069	-0.019	-0.038	-0.011
Alkylgreater than5 C	0.041	0.007	-0.084	-0.057	-0.141	-0.006	0.300
Phosphonate	-0.272	0.210	0.111	-0.057	-0.045	-0.022	0.013
Hydrozone	-0.273	0.210	0.162	0.011	0.008	0.037	0.025
Other_1	-0.361	0.228	0.138	-0.059	-0.048	-0.048	-0.011
Phosphate	-0.361	0.228	0.138	-0.059	-0.048	-0.048	-0.011
Carbamate	0.020	0.009	-0.008	0.007	0.008	-0.038	-0.023
Nitro	0.077	0.114	0.125	0.065	0.043	0.452	-0.030
Nitrate	0.033	-0.009	0.015	0.048	0.092	0.035	-0.309
Steroid	-0.266	-0.059	-0.268	0.135	0.074	0.051	-0.005
Hormone	0.010	0.057	-0.073	-0.033	-0.016	-0.125	-0.163
O-heterocyclic	0.082	0.063	0.040	0.046	0.051	0.085	-0.306
N-heterocyclic	0.107	0.183	-0.012	-0.127	-0.061	0.131	0.059

S-heterocyclic	0.057	0.094	0.097	0.040	0.024	0.379	0.016
Long alkyl	0.028	0.007	0.026	-0.037	0.007	-0.107	-0.015
Phenyl ring	0.156	0.186	-0.057	-0.166	-0.122	-0.129	-0.214
Erythromycin deriv	-0.035	-0.203	0.182	-0.317	0.118	0.041	0.023
Tetracycline	-0.028	-0.189	0.174	0.142	-0.421	0.017	-0.073
Macrocyclic	0.042	0.036	-0.063	-0.057	-0.095	-0.032	0.358
Macrolide	-0.045	-0.239	0.158	-0.320	0.131	0.044	0.018
Benzodiazepine	0.013	0.001	0.044	0.071	0.053	0.033	0.062
Barbiturate	0.014	0.024	-0.017	-0.007	-0.000	0.011	0.039
Water	-0.011	-0.085	-0.031	0.029	-0.234	0.059	-0.042
Ethanol	-0.002	-0.099	0.104	0.092	-0.199	0.001	-0.030
HCl	-0.021	-0.156	0.152	0.126	-0.356	0.013	-0.066
Na+	-0.336	0.216	0.132	-0.049	-0.040	-0.054	-0.007
Gd3+	0.076	0.014	0.253	0.222	0.175	-0.195	0.138

Variable	PC8	PC9	PC10	PC11	PC12	PC13	PC14
Primary	0.126	-0.122	0.085	-0.115	-0.004	0.111	0.026
Secondary	0.047	-0.048	-0.029	0.066	-0.139	-0.015	0.111
Tertiary	0.120	-0.100	0.115	-0.061	0.055	0.090	-0.052
Aromatic/enamine	0.089	-0.089	0.104	0.146	-0.008	-0.043	-0.141
Primary_1	-0.050	-0.030	0.079	0.256	0.081	-0.133	-0.112
Secondary_1	-0.273	-0.109	0.134	-0.000	-0.043	0.031	-0.213
Tertiary_1	0.011	-0.017	-0.096	0.050	0.036	-0.116	-0.056
Vinyl alcohol	0.021	-0.008	0.028	-0.008	0.054	0.012	0.013
Phenol	0.121	-0.094	-0.129	0.058	0.050	-0.204	-0.261
Carboxylic	0.113	-0.111	0.112	0.038	-0.061	0.110	-0.110
Sulfonated	0.000	0.000	0.000	0.000	0.000	0.000	0.000
Other	0.000	0.000	0.000	0.000	0.000	0.000	0.000
Ketone	0.092	-0.126	0.075	0.176	-0.003	0.027	-0.021
Aldehyde	0.000	0.000	0.000	-0.000	-0.000	-0.000	-0.000
Enone	0.057	-0.017	-0.212	0.090	0.219	0.244	-0.069
Ester	0.035	-0.102	0.137	-0.141	-0.306	0.078	-0.152
1o amide	-0.136	-0.044	-0.070	-0.252	-0.315	0.152	-0.349
2o amide	-0.084	-0.135	0.145	0.155	0.167	-0.104	-0.093
3o amide	-0.273	-0.107	-0.094	0.062	0.197	0.087	0.195
Anhydride	-0.000	-0.000	-0.000	0.000	0.000	0.000	0.000
Epoxide	-0.000	-0.000	-0.000	0.000	0.000	0.000	0.000
Thioester	0.094	-0.030	0.095	-0.454	0.369	-0.171	-0.072
Oxime	0.014	-0.066	0.001	0.047	0.156	0.094	-0.168
Oxazolidinone	-0.001	0.049	-0.052	0.034	-0.002	-0.036	0.144
Urea	0.004	0.588	0.079	0.111	0.082	0.171	-0.178
Guanidine	0.067	-0.120	0.323	0.109	0.069	-0.084	-0.077
Ether	0.004	-0.024	0.132	-0.034	-0.073	0.022	0.192
Sulfonamide	0.002	0.034	-0.050	-0.034	-0.063	0.028	0.251
Sulfone	0.014	0.047	0.107	-0.128	-0.207	0.013	0.223
N-Oxide	-0.447	-0.018	0.111	0.025	0.113	-0.048	-0.118
Nitrile	0.000	0.000	0.000	-0.000	-0.000	-0.000	-0.000
Thiol	0.000	0.000	0.000	-0.000	-0.000	-0.000	-0.000
Thioether	0.053	-0.061	-0.163	-0.040	0.004	0.034	-0.001
Fluorine	0.078	-0.094	0.069	-0.132	0.211	-0.122	-0.106
Pyridine	-0.056	0.018	0.094	-0.105	-0.183	0.024	-0.042
Alkyl halide	0.091	-0.121	0.111	0.261	-0.067	0.027	-0.057

Aryl halide	0.119	-0.089	-0.201	-0.127	0.038	0.219	-0.101
Alkene	0.020	0.080	-0.260	-0.041	-0.123	-0.351	-0.132
Alkylgreater than5 C	-0.252	-0.096	-0.121	-0.215	-0.229	0.144	-0.277
Phosphonate	-0.048	0.115	0.045	0.075	-0.017	0.019	-0.016
Hydrozone	-0.015	-0.036	0.019	-0.135	0.060	-0.021	0.008
Other_1	-0.062	-0.036	0.016	0.027	-0.038	0.008	0.022
Phosphate	-0.062	-0.036	0.016	0.027	-0.038	0.008	0.022
Carbamate	0.000	0.034	-0.041	-0.012	-0.022	-0.012	0.120
Nitro	0.035	-0.063	-0.204	0.106	-0.082	0.068	0.006
Nitrate	-0.444	-0.018	0.111	0.024	0.112	-0.048	-0.117
Steroid	0.112	-0.149	0.022	0.151	0.130	0.061	-0.126
Hormone	0.141	-0.087	-0.286	-0.025	0.219	0.352	-0.150
O-heterocyclic	-0.241	-0.059	-0.026	-0.003	-0.001	0.061	0.108
N-heterocyclic	0.117	0.138	0.287	-0.176	-0.053	-0.001	-0.025
S-heterocyclic	0.080	-0.033	-0.180	0.091	-0.080	0.052	0.007
Long alkyl	0.044	0.058	-0.283	0.083	-0.064	-0.500	-0.198
Phenyl ring	0.126	-0.122	0.085	-0.115	-0.004	0.111	0.026
Erythromycin deriv	0.014	-0.067	0.001	0.048	0.158	0.096	-0.170
Tetracycline	0.022	0.001	0.025	0.003	0.061	-0.001	-0.005
Macrocyclic	-0.272	-0.102	-0.163	0.011	0.180	0.125	0.198
Macrolide	0.021	-0.039	0.022	0.020	0.033	0.050	0.039
Benzodiazepine	0.050	0.035	0.098	-0.376	0.264	-0.118	-0.005
Barbiturate	0.007	0.559	0.065	0.085	0.085	0.182	-0.177
Water	0.073	-0.091	0.229	0.169	0.000	0.001	0.005
Ethanol	0.011	0.043	-0.017	-0.023	0.063	-0.013	0.111
HCl	0.011	0.009	0.011	-0.013	0.070	-0.003	0.010
Na+	-0.061	-0.022	0.004	0.029	-0.041	0.004	0.044
Gd3+	0.076	-0.119	0.146	0.128	-0.093	0.127	-0.122

Variable	PC15	PC16	PC17	PC18	PC19	PC20	PC21
Primary	-0.208	0.067	0.025	-0.067	-0.003	-0.056	0.025
Secondary	-0.226	0.017	0.075	0.221	-0.159	0.132	-0.124
Tertiary	0.057	0.130	-0.088	-0.010	0.053	0.021	0.097
Aromatic/enamine	0.031	-0.015	-0.009	0.042	-0.026	-0.134	0.045
Primary_1	0.121	-0.281	0.277	-0.141	0.035	0.011	0.333
Secondary_1	0.012	-0.173	0.011	-0.092	-0.042	0.150	0.013
Tertiary_1	-0.051	0.056	-0.074	0.008	-0.069	-0.148	0.081
Vinyl alcohol	-0.011	0.052	0.067	-0.039	-0.034	-0.005	-0.029
Phenol	0.074	0.142	0.305	0.141	0.163	0.350	-0.187
Carboxylic	0.067	0.194	-0.030	-0.052	0.028	-0.137	-0.144
Sulfonated	0.000	0.000	0.000	0.000	0.000	0.000	0.000
Other	0.000	0.000	0.000	0.000	0.000	0.000	0.000
Ketone	-0.082	0.105	0.024	-0.044	0.160	-0.000	0.039
Aldehyde	-0.000	0.000	0.000	-0.000	-0.000	-0.000	0.000
Enone	0.328	-0.296	-0.214	0.217	-0.289	-0.059	-0.175
Ester	0.063	0.117	0.112	-0.211	-0.043	0.028	0.007
1o amide	-0.085	-0.138	0.045	0.304	0.081	-0.026	0.055
2o amide	0.032	-0.140	0.195	-0.073	-0.015	-0.057	0.091
3o amide	0.011	0.234	-0.029	-0.073	-0.031	0.049	-0.039
Anhydride	0.000	-0.000	-0.000	0.000	0.000	0.000	-0.000
Epoxide	0.000	-0.000	-0.000	0.000	0.000	0.000	-0.000
Thioester	-0.060	0.057	0.086	0.108	-0.005	-0.059	-0.050

Oxime	-0.203	-0.083	-0.033	-0.119	-0.029	0.113	-0.186
Oxazolidinone	-0.032	-0.192	-0.123	-0.035	0.764	-0.208	-0.344
Urea	-0.079	0.144	0.081	0.049	0.011	-0.008	0.038
Guanidine	0.058	0.004	-0.188	0.269	-0.032	-0.136	-0.086
Ether	0.274	0.148	0.160	0.244	0.045	-0.035	0.068
Sulfonamide	-0.182	-0.012	-0.000	0.165	-0.194	0.095	-0.026
Sulfone	0.419	0.138	0.298	0.077	-0.048	0.001	-0.096
N-Oxide	0.127	0.124	-0.032	-0.001	0.052	-0.045	-0.051
Nitrile	-0.000	0.000	0.000	-0.000	0.000	0.000	0.000
Thiol	-0.000	0.000	0.000	-0.000	-0.000	0.000	-0.000
Thioether	-0.022	0.072	0.060	0.039	0.022	0.020	-0.085
Fluorine	-0.169	0.075	0.322	0.022	-0.132	-0.332	-0.145
Pyridine	0.107	-0.095	-0.108	-0.367	-0.154	0.257	-0.286
Alkyl halide	-0.097	0.209	-0.100	-0.052	0.097	0.288	0.129
Aryl halide	0.077	0.244	0.058	-0.217	0.164	-0.024	0.218
Alkene	0.043	0.182	-0.347	-0.225	-0.153	-0.371	0.177
Alkylgreater than5 C	-0.070	0.001	-0.029	0.176	0.049	-0.053	0.007
Phosphonate	0.012	-0.043	0.020	0.040	0.010	-0.008	0.135
Hydrozone	0.009	0.056	0.000	0.093	0.037	-0.015	-0.133
Other_1	-0.002	0.003	0.002	0.056	0.013	0.019	0.043
Phosphate	-0.002	0.003	0.002	0.056	0.013	0.019	0.043
Carbamate	-0.099	-0.128	-0.033	0.038	-0.039	0.038	-0.008
Nitro	-0.037	0.033	0.074	-0.006	-0.001	0.052	0.062
Nitrate	0.126	0.123	-0.032	-0.001	0.052	-0.045	-0.051
Steroid	-0.007	0.033	0.028	0.054	-0.141	-0.030	-0.058
Hormone	0.308	0.010	-0.029	0.009	0.044	0.012	0.034
O-heterocyclic	-0.237	0.103	0.066	0.016	-0.103	-0.030	0.068
N-heterocyclic	0.158	0.025	-0.150	-0.212	-0.099	0.048	-0.210
S-heterocyclic	0.032	0.039	0.044	-0.028	0.035	0.057	0.008
Long alkyl	0.089	0.202	-0.008	0.097	-0.030	0.183	-0.189
Phenyl ring	-0.208	0.067	0.025	-0.067	-0.003	-0.056	0.025
Erythromycin deriv	-0.206	-0.084	-0.034	-0.121	-0.030	0.115	-0.189
Tetracycline	0.032	0.041	0.057	-0.020	-0.028	0.008	-0.044
Macrocyclic	-0.034	0.318	-0.025	-0.070	-0.024	0.057	-0.105
Macrolide	0.035	0.030	0.066	0.106	0.082	-0.068	0.182
Benzodiazepine	0.049	-0.028	-0.245	0.054	0.141	0.390	0.372
Barbiturate	-0.098	0.193	0.078	0.054	0.011	0.006	-0.031
Water	0.022	0.200	-0.335	0.247	0.019	0.124	-0.014
Ethanol	0.092	-0.088	0.132	-0.176	-0.002	0.018	-0.026
HCl	0.054	-0.011	0.156	-0.135	-0.045	0.008	-0.074
Na+	0.000	-0.018	-0.013	0.060	0.021	0.010	0.043
Gd3+	0.062	0.161	-0.046	-0.030	0.002	-0.182	-0.034

Variable	PC22	PC23	PC24	PC25	PC26	PC27	PC28
Primary	0.076	-0.082	-0.000	-0.159	0.179	-0.006	0.076
Secondary	-0.102	0.062	-0.081	-0.073	-0.100	0.033	-0.224
Tertiary	-0.174	-0.022	-0.025	-0.035	0.083	0.052	0.013
Aromatic/enamine	-0.194	0.050	0.004	-0.027	-0.017	0.098	-0.128
Primary_1	0.192	0.226	0.062	0.148	-0.101	-0.040	0.229
Secondary_1	0.365	-0.185	-0.077	-0.063	-0.057	0.413	-0.352
Tertiary_1	0.005	0.036	-0.162	-0.021	0.061	-0.012	-0.138
Vinyl alcohol	0.102	0.207	-0.101	-0.064	0.082	-0.053	0.054

Phenol	-0.078	-0.057	-0.105	-0.004	0.040	0.005	-0.024
Carboxylic	0.023	-0.009	0.040	0.062	0.079	-0.117	0.057
Sulfonated	0.000	0.000	0.000	0.000	0.000	0.000	0.000
Other	0.000	0.000	0.000	0.000	0.000	0.000	0.000
Ketone	-0.101	0.035	-0.100	0.003	0.188	-0.076	-0.066
Aldehyde	-0.000	-0.000	-0.000	0.000	0.000	-0.000	0.000
Enone	-0.058	0.066	-0.025	-0.060	0.274	0.058	0.028
Ester	-0.009	0.063	-0.048	0.022	0.036	0.107	-0.000
1o amide	-0.045	0.026	0.080	0.022	-0.027	-0.076	0.094
2o amide	-0.111	0.132	-0.006	0.088	0.164	-0.092	-0.253
3o amide	0.093	-0.012	-0.033	-0.030	-0.055	0.083	0.053
Anhydride	0.000	0.000	0.000	-0.000	-0.000	0.000	-0.000
Epoxide	0.000	0.000	0.000	-0.000	-0.000	0.000	-0.000
Thioester	0.113	0.024	0.019	0.097	-0.056	-0.005	0.027
Oxime	-0.010	0.122	0.279	-0.069	-0.069	-0.009	0.034
Oxazolidinone	-0.058	0.191	-0.033	-0.017	0.013	0.193	-0.074
Urea	0.003	0.009	-0.003	-0.022	0.007	0.066	0.040
Guanidine	-0.045	-0.108	0.019	0.047	-0.154	0.193	-0.108
Ether	0.007	-0.094	-0.100	0.056	-0.008	0.038	0.008
Sulfonamide	-0.286	0.305	0.010	0.406	-0.258	-0.001	-0.239
Sulfone	0.090	0.283	0.403	-0.276	0.054	0.119	-0.121
N-Oxide	-0.129	0.005	0.003	-0.032	-0.071	-0.217	-0.006
Nitrile	-0.000	0.000	0.000	0.000	-0.000	0.000	-0.000
Thiol	-0.000	-0.000	-0.000	0.000	-0.000	0.000	0.000
Thioether	0.243	-0.051	0.085	0.225	0.132	-0.049	-0.036
Fluorine	-0.084	0.021	-0.128	-0.137	-0.276	0.117	0.100
Pyridine	-0.031	0.037	-0.366	0.105	-0.093	0.128	0.115
Alkyl halide	-0.053	-0.015	0.219	0.105	0.123	-0.156	-0.148
Aryl halide	0.022	0.081	0.043	0.158	-0.342	0.056	-0.129
Alkene	0.021	0.064	0.156	-0.008	0.012	0.086	-0.186
Alkylgreater than5 C	-0.146	-0.012	0.046	0.034	0.039	-0.043	-0.003
Phosphonate	-0.093	0.072	0.071	-0.106	-0.064	-0.000	0.291
Hydrozone	0.278	-0.089	0.173	0.327	0.187	-0.191	-0.122
Other_1	-0.037	-0.006	0.020	-0.085	0.015	0.057	-0.066
Phosphate	-0.037	-0.006	0.020	-0.085	0.015	0.057	-0.066
Carbamate	0.168	-0.202	0.042	-0.338	-0.161	-0.350	-0.145
Nitro	-0.012	0.023	-0.056	-0.116	-0.017	0.173	0.031
Nitrate	-0.128	0.005	0.003	-0.032	-0.070	-0.216	-0.006
Steroid	-0.101	0.020	0.021	-0.069	0.068	0.043	-0.072
Hormone	0.016	0.051	-0.035	0.082	-0.128	0.008	-0.022
O-heterocyclic	-0.087	0.241	0.048	0.122	0.311	0.284	0.259
N-heterocyclic	-0.135	0.107	-0.055	0.079	0.055	-0.251	0.028
S-heterocyclic	-0.113	-0.015	-0.113	-0.255	-0.243	-0.228	0.018
Long alkyl	-0.008	-0.007	-0.000	-0.026	0.109	0.085	0.181
Phenyl ring	0.076	-0.082	-0.000	-0.159	0.179	-0.006	0.076
Erythromycin deriv	-0.010	0.123	0.283	-0.070	-0.070	-0.009	0.035
Tetracycline	0.094	0.187	-0.102	-0.047	-0.006	-0.081	-0.033
Macrocyclic	-0.012	-0.029	-0.055	-0.096	0.041	0.047	-0.011
Macrolide	-0.023	-0.168	-0.368	0.153	0.129	-0.042	0.030
Benzodiazepine	-0.198	0.139	-0.016	-0.235	0.096	0.128	-0.081
Barbiturate	0.050	-0.044	-0.043	0.007	0.060	0.103	-0.194
Water	0.258	0.028	0.082	0.083	-0.314	0.134	0.267

Ethanol	-0.392	-0.570	0.355	0.161	-0.051	0.195	0.139
HCl	-0.083	-0.053	0.023	-0.002	0.045	0.032	-0.269
Na+	-0.029	-0.024	0.040	-0.095	0.002	0.025	-0.034
Gd3+	-0.088	0.013	-0.019	-0.135	0.013	0.057	0.027

Variable	PC29	PC30	PC31	PC32	PC33	PC34	PC35
Primary	-0.185	-0.056	-0.166	-0.072	0.082	0.052	0.132
Secondary	-0.047	0.007	-0.337	0.317	-0.201	0.144	-0.287
Tertiary	0.067	-0.061	-0.035	0.039	0.227	0.100	0.147
Aromatic/enamine	0.217	0.193	0.017	0.078	0.049	-0.107	-0.074
Primary_1	-0.061	-0.220	-0.198	-0.174	0.006	0.172	0.021
Secondary_1	-0.165	0.258	0.045	-0.011	0.162	0.026	0.210
Tertiary_1	-0.020	-0.044	-0.109	0.238	-0.132	0.088	0.008
Vinyl alcohol	0.025	-0.006	0.061	0.107	0.165	-0.058	0.120
Phenol	0.065	0.058	0.041	0.112	0.133	-0.066	-0.128
Carboxylic	0.012	0.068	0.177	-0.137	-0.047	-0.168	-0.029
Sulfonated	0.000	0.000	0.000	0.000	0.000	0.000	0.000
Other	0.000	0.000	0.000	0.000	0.000	0.000	0.000
Ketone	0.007	0.038	0.263	0.068	-0.094	0.554	0.108
Aldehyde	-0.000	-0.000	0.000	-0.000	-0.000	0.000	-0.000
Enone	-0.048	-0.034	-0.168	-0.089	0.232	-0.070	-0.122
Ester	0.086	0.006	-0.066	-0.065	0.346	0.155	-0.360
1o amide	-0.044	-0.052	-0.057	-0.068	-0.042	-0.064	0.073
2o amide	0.086	-0.153	-0.042	0.179	0.105	-0.073	-0.022
3o amide	-0.051	0.026	-0.024	-0.108	0.013	0.063	-0.002
Anhydride	0.000	0.000	-0.000	-0.000	-0.000	0.000	-0.000
Epoxide	0.000	0.000	-0.000	0.000	0.000	-0.000	0.000
Thioester	0.022	-0.009	-0.187	0.005	-0.019	-0.039	-0.007
Oxime	-0.012	-0.003	0.062	-0.029	-0.014	0.022	-0.002
Oxazolidinone	0.019	-0.027	-0.222	-0.051	0.110	-0.029	0.016
Urea	0.038	0.138	-0.137	0.020	0.032	0.055	0.030
Guanidine	0.094	-0.065	0.078	-0.128	-0.172	0.203	0.092
Ether	-0.026	-0.004	-0.101	-0.127	-0.046	-0.075	0.128
Sulfonamide	-0.078	0.054	0.042	-0.223	0.324	0.072	0.286
Sulfone	0.022	-0.073	-0.044	0.120	-0.058	0.058	0.125
N-Oxide	-0.041	-0.041	-0.062	0.028	0.021	0.022	-0.021
Nitrile	0.000	-0.000	0.000	0.000	0.000	-0.000	0.000
Thiol	-0.000	-0.000	-0.000	0.000	0.000	0.000	0.000
Thioether	0.061	0.016	-0.084	0.049	-0.032	0.066	0.008
Fluorine	0.007	-0.044	-0.028	-0.024	-0.051	-0.121	0.166
Pyridine	0.196	-0.236	-0.243	0.070	-0.177	-0.032	0.181
Alkyl halide	0.018	0.003	-0.343	-0.045	-0.153	-0.381	0.291
Aryl halide	0.116	0.030	0.036	-0.018	0.102	-0.076	-0.201
Alkene	-0.023	-0.050	-0.180	0.029	0.050	0.008	0.044
Alkylgreater than5 C	0.054	-0.049	-0.039	0.010	-0.073	0.074	0.076
Phosphonate	0.056	0.612	-0.257	0.054	0.038	0.086	0.064
Hydrozone	-0.021	-0.020	-0.111	0.044	0.079	0.189	-0.054
Other_1	-0.041	-0.121	0.030	-0.058	0.075	0.020	0.025
Phosphate	-0.041	-0.121	0.030	-0.058	0.075	0.020	0.025
Carbamate	0.663	-0.081	-0.078	-0.182	0.209	0.109	0.089
Nitro	0.065	-0.047	0.031	-0.026	-0.065	-0.095	0.082
Nitrate	-0.041	-0.041	-0.062	0.028	0.021	0.022	-0.021

Steroid	-0.028	-0.023	-0.138	-0.054	0.146	-0.030	-0.053
Hormone	0.109	-0.007	0.006	0.084	-0.269	0.181	0.254
O-heterocyclic	0.309	-0.127	0.012	-0.101	-0.088	0.053	-0.092
N-heterocyclic	-0.091	0.133	0.072	-0.052	-0.003	0.030	0.001
S-heterocyclic	-0.394	-0.186	-0.130	-0.102	0.090	0.164	-0.040
Long alkyl	0.016	0.015	0.053	-0.270	0.084	0.112	0.179
Phenyl ring	-0.185	-0.056	-0.166	-0.072	0.082	0.052	0.132
Erythromycin deriv	-0.012	-0.003	0.062	-0.030	-0.014	0.022	-0.002
Tetracycline	0.015	0.031	0.054	0.143	0.137	-0.135	0.121
Macrocyclic	0.007	0.096	-0.052	0.024	0.069	-0.009	0.002
Macrolide	0.065	0.024	-0.128	-0.114	0.113	-0.087	-0.005
Benzodiazepine	0.011	-0.059	0.010	-0.043	-0.043	0.053	-0.057
Barbiturate	-0.018	-0.340	0.043	-0.046	0.046	0.011	0.001
Water	0.001	-0.147	-0.107	0.056	0.101	0.025	-0.225
Ethanol	0.002	-0.130	-0.076	0.180	0.140	0.068	0.050
HCl	-0.008	0.154	-0.184	-0.576	-0.342	0.072	-0.279
Na+	-0.066	-0.152	0.256	-0.104	-0.064	-0.329	-0.125
Gd3+	0.078	-0.044	-0.048	-0.070	0.015	0.039	0.056

Variable	PC36	PC37	PC38	PC39	PC40	PC41	PC42
Primary	-0.041	-0.010	0.098	0.075	-0.032	0.098	-0.058
Secondary	-0.105	-0.200	-0.147	0.027	0.260	-0.004	-0.077
Tertiary	-0.010	-0.115	0.162	0.218	0.109	-0.030	-0.023
Aromatic/enamine	-0.008	-0.036	-0.038	0.016	0.110	0.181	0.037
Primary_1	-0.043	-0.172	0.110	-0.042	0.028	-0.084	-0.207
Secondary_1	-0.120	0.132	-0.161	-0.103	0.032	-0.059	-0.112
Tertiary_1	-0.164	-0.053	-0.096	-0.057	-0.456	-0.028	-0.306
Vinyl alcohol	0.188	-0.038	-0.109	0.009	0.089	-0.122	0.009
Phenol	0.087	0.159	0.329	-0.113	-0.079	-0.391	-0.063
Carboxylic	-0.074	0.176	-0.158	-0.169	-0.246	0.014	-0.112
Sulfonated	0.000	0.000	0.000	0.000	0.000	0.000	0.000
Other	0.000	0.000	0.000	0.000	0.000	0.000	0.000
Ketone	-0.074	0.121	0.142	-0.197	0.277	0.169	-0.188
Aldehyde	0.000	-0.000	0.000	0.000	0.000	-0.000	0.000
Enone	0.031	-0.011	0.073	-0.238	0.208	0.057	-0.175
Ester	0.070	-0.107	-0.130	0.355	0.033	-0.153	0.004
1o amide	-0.077	-0.190	-0.102	-0.034	0.004	0.114	0.023
2o amide	-0.144	0.270	-0.186	0.104	-0.028	0.220	0.193
3o amide	-0.015	-0.211	-0.064	-0.037	0.020	-0.200	-0.102
Anhydride	0.000	-0.000	0.000	-0.000	-0.000	0.000	0.000
Epoxide	-0.000	0.000	-0.000	-0.000	-0.000	0.000	-0.000
Thioester	0.055	0.174	-0.073	0.012	0.215	0.023	-0.025
Oxime	0.037	0.005	0.031	-0.010	0.075	0.013	0.049
Oxazolidinone	-0.068	-0.032	-0.061	0.025	-0.030	-0.050	-0.062
Urea	0.060	0.069	0.000	0.052	-0.030	0.045	-0.105
Guanidine	0.434	-0.208	-0.093	0.018	-0.039	-0.110	-0.139
Ether	0.037	-0.047	0.054	-0.127	0.041	-0.040	0.156
Sulfonamide	-0.037	0.119	0.126	0.090	-0.114	0.015	-0.083
Sulfone	-0.030	0.112	0.023	-0.047	-0.038	0.121	-0.149
N-Oxide	-0.017	-0.074	0.092	0.068	0.004	0.061	-0.020
Nitrile	-0.000	-0.000	0.000	-0.000	0.000	0.000	-0.000
Thiol	-0.000	-0.000	0.000	0.000	0.000	0.000	-0.000

Thioether	-0.006	-0.115	0.155	0.081	-0.082	0.067	-0.104
Fluorine	-0.172	-0.032	0.019	-0.085	0.068	-0.069	-0.062
Pyridine	0.176	0.101	0.129	-0.048	-0.049	0.211	-0.014
Alkyl halide	0.051	-0.044	-0.136	-0.065	0.156	-0.123	-0.079
Aryl halide	0.081	-0.181	-0.083	-0.354	0.056	0.286	-0.130
Alkene	0.031	0.141	0.180	-0.021	0.156	-0.224	0.023
Alkylgreater than5 C	-0.025	0.203	0.147	-0.110	-0.021	-0.078	-0.115
Phosphonate	0.070	0.108	-0.012	0.038	-0.044	-0.003	-0.022
Hydrozone	0.134	-0.027	-0.060	-0.006	-0.055	0.005	0.059
Other_1	-0.120	-0.081	-0.020	-0.117	0.062	-0.118	0.290
Phosphate	-0.120	-0.081	-0.020	-0.117	0.062	-0.118	0.290
Carbamate	-0.098	0.033	-0.048	-0.046	-0.026	-0.068	-0.119
Nitro	-0.227	-0.169	0.234	0.020	0.001	-0.008	0.022
Nitrate	-0.016	-0.074	0.092	0.067	0.004	0.061	-0.020
Steroid	-0.024	-0.162	-0.004	0.037	-0.483	0.035	0.011
Hormone	-0.177	0.090	-0.204	0.405	0.015	-0.215	0.126
O-heterocyclic	0.029	0.164	-0.286	-0.159	-0.025	-0.229	-0.036
N-heterocyclic	-0.429	-0.185	-0.139	-0.160	0.089	-0.277	-0.149
S-heterocyclic	0.246	0.317	-0.337	-0.104	-0.027	-0.082	-0.024
Long alkyl	-0.114	-0.137	-0.311	0.106	0.067	0.334	0.007
Phenyl ring	-0.041	-0.010	0.098	0.075	-0.032	0.098	-0.058
Erythromycin deriv	0.038	0.005	0.032	-0.010	0.076	0.013	0.050
Tetracycline	0.092	-0.119	-0.034	-0.029	0.053	-0.013	0.006
Macrocyclic	0.014	0.005	0.031	0.049	0.018	0.145	0.001
Macrolide	-0.096	0.118	-0.083	-0.009	0.178	0.047	-0.092
Benzodiazepine	0.008	-0.035	-0.013	-0.067	-0.175	0.043	-0.026
Barbiturate	-0.043	-0.041	-0.001	-0.028	0.029	-0.007	0.044
Water	-0.359	0.295	0.094	0.009	-0.017	0.083	0.037
Ethanol	-0.083	0.008	-0.110	-0.064	-0.009	-0.017	-0.118
HCl	-0.087	0.120	0.145	0.065	-0.014	0.038	0.085
Na+	0.021	0.118	0.001	0.405	0.174	0.011	-0.527
Gd3+	0.108	0.059	0.070	-0.009	0.009	0.030	0.149

Variable	PC43	PC44	PC45	PC46	PC47	PC48	PC49
Primary	-0.044	0.008	0.054	-0.010	-0.025	-0.027	-0.001
Secondary	-0.168	-0.077	-0.035	0.057	0.038	-0.106	0.116
Tertiary	0.432	0.049	-0.309	0.126	-0.047	0.076	0.134
Aromatic/enamine	-0.146	-0.057	0.237	0.227	-0.087	0.309	0.083
Primary_1	-0.112	0.018	0.111	0.216	0.106	0.050	-0.037
Secondary_1	-0.020	0.002	-0.072	0.022	0.070	0.055	-0.007
Tertiary_1	0.162	0.260	0.216	-0.143	-0.015	0.178	-0.152
Vinyl alcohol	-0.144	-0.003	0.152	-0.061	0.119	0.274	0.518
Phenol	0.031	-0.029	-0.039	0.008	-0.105	0.117	-0.113
Carboxylic	-0.229	-0.125	0.267	0.154	-0.244	-0.203	0.098
Sulfonated	0.000	0.000	0.000	0.000	0.000	0.000	0.000
Other	0.000	0.000	0.000	0.000	0.000	0.000	0.000
Ketone	-0.170	-0.013	-0.035	-0.030	-0.123	0.043	-0.059
Aldehyde	0.000	0.000	0.000	-0.000	-0.000	0.000	0.000
Enone	0.078	-0.003	0.146	-0.143	0.001	-0.008	-0.068
Ester	-0.036	0.018	0.286	-0.115	-0.006	-0.110	-0.108
1o amide	-0.060	-0.157	-0.159	0.010	-0.334	0.304	-0.251

2o amide	0.023	-0.033	-0.188	-0.298	-0.022	-0.125	-0.016
3o amide	-0.061	-0.059	-0.090	-0.337	-0.450	-0.013	0.178
Anhydride	0.000	-0.000	-0.000	-0.000	-0.000	0.000	0.000
Epoxide	-0.000	-0.000	0.000	-0.000	-0.000	0.000	-0.000
Thioester	-0.169	0.541	0.019	0.061	-0.197	0.016	-0.054
Oxime	-0.063	-0.038	-0.026	0.007	0.004	-0.023	0.014
Oxazolidinone	-0.043	-0.005	-0.006	0.002	0.023	-0.007	0.008
Urea	-0.041	-0.033	-0.035	-0.057	-0.012	-0.031	0.002
Guanidine	0.102	0.028	0.016	0.064	-0.014	-0.187	-0.058
Ether	-0.416	0.004	-0.155	-0.163	0.236	0.116	-0.006
Sulfonamide	-0.056	0.010	0.123	-0.048	-0.062	0.030	-0.051
Sulfone	0.172	-0.028	0.003	0.118	-0.151	-0.033	0.052
N-Oxide	-0.015	-0.002	0.063	-0.008	-0.040	-0.032	0.028
Nitrile	-0.000	-0.000	0.000	-0.000	0.000	-0.000	-0.000
Thiol	-0.000	0.000	-0.000	-0.000	-0.000	0.000	-0.000
Thioether	-0.063	-0.158	-0.003	-0.152	0.087	-0.075	-0.012
Fluorine	0.213	-0.419	0.109	-0.122	0.160	0.003	-0.007
Pyridine	-0.060	-0.064	-0.037	0.024	-0.127	0.035	0.035
Alkyl halide	0.143	0.037	0.145	-0.059	0.021	0.013	-0.042
Aryl halide	0.111	0.089	-0.142	-0.081	0.097	-0.022	0.065
Alkene	-0.194	-0.181	-0.144	0.083	-0.103	-0.008	-0.017
Alkylgreater than5 C	0.016	0.161	0.060	0.027	0.364	-0.301	0.326
Phosphonate	0.002	0.003	-0.049	-0.099	-0.031	-0.110	-0.005
Hydrozone	0.062	-0.284	0.014	-0.025	0.126	0.083	-0.007
Other_1	0.035	0.103	0.129	0.079	-0.057	-0.058	0.006
Phosphate	0.035	0.103	0.129	0.079	-0.057	-0.058	0.006
Carbamate	-0.071	0.003	-0.023	-0.016	0.000	0.027	-0.026
Nitro	-0.113	0.167	0.074	-0.102	-0.062	-0.215	-0.003
Nitrate	-0.015	-0.002	0.062	-0.008	-0.040	-0.031	0.027
Steroid	-0.247	0.058	-0.421	0.252	0.021	-0.013	0.198
Hormone	-0.082	-0.059	0.056	0.065	0.008	0.010	-0.019
O-heterocyclic	0.054	-0.008	-0.180	0.053	0.050	0.054	-0.113
N-heterocyclic	-0.009	0.059	-0.195	0.025	0.180	0.071	-0.060
S-heterocyclic	0.041	-0.029	-0.074	0.031	0.009	0.031	-0.043
Long alkyl	0.029	0.061	0.004	-0.057	0.109	-0.071	0.074
Phenyl ring	-0.044	0.008	0.054	-0.010	-0.025	-0.027	-0.001
Erythromycin deriv	-0.064	-0.038	-0.027	0.007	0.004	-0.023	0.015
Tetracycline	-0.129	-0.047	-0.112	0.102	0.028	-0.496	-0.404
Macrocyclic	-0.026	-0.017	0.113	0.409	0.266	0.198	-0.318
Macrolide	0.091	-0.188	-0.023	0.236	-0.157	-0.195	0.106
Benzodiazepine	-0.150	-0.326	0.145	-0.090	0.083	-0.056	-0.012
Barbiturate	-0.029	0.012	0.061	0.045	-0.021	0.014	0.009
Water	0.041	-0.053	-0.011	-0.025	0.028	0.022	0.012
Ethanol	-0.156	-0.019	0.068	-0.018	-0.002	-0.025	-0.029
HCl	0.113	0.002	-0.022	-0.002	-0.001	0.023	0.023
Na+	-0.156	-0.000	-0.156	-0.069	0.097	0.069	-0.033
Gd3+	-0.110	0.105	-0.012	-0.359	0.201	0.153	-0.249

Variable	PC50	PC51	PC52	PC53	PC54	PC55	PC56
Primary	0.047	0.011	-0.022	0.096	0.674	0.172	-0.034
Secondary	-0.042	0.028	0.000	-0.000	0.000	0.000	-0.000
Tertiary	0.073	0.349	-0.000	0.000	0.000	0.000	0.000

Aromatic/enamine	0.477	-0.018	-0.000	0.000	-0.000	-0.000	-0.000
Primary_1	0.034	0.013	-0.000	0.000	0.000	-0.000	-0.000
Secondary_1	0.005	0.010	-0.000	0.000	0.000	-0.000	0.000
Tertiary_1	0.058	0.072	-0.000	0.000	0.000	0.000	0.000
Vinyl alcohol	-0.301	-0.096	0.000	-0.000	-0.000	-0.000	0.000
Phenol	-0.033	-0.041	-0.000	-0.000	-0.000	-0.000	0.000
Carboxylic	-0.127	0.274	-0.000	-0.000	0.000	0.000	0.000
Sulfonated	0.000	0.000	-0.000	0.000	-0.000	0.000	0.000
Other	0.000	0.000	0.000	0.000	0.000	0.000	-0.000
Ketone	-0.028	0.072	-0.000	0.000	0.000	0.000	0.000
Aldehyde	0.000	0.000	-0.001	0.002	-0.003	0.010	-0.035
Enone	-0.033	0.114	-0.000	-0.000	0.000	0.000	0.000
Ester	-0.042	0.118	-0.000	-0.000	0.000	0.000	-0.000
1o amide	-0.179	0.006	-0.000	-0.000	-0.000	-0.000	0.000
2o amide	-0.236	0.134	-0.000	-0.000	0.000	0.000	0.000
3o amide	0.270	-0.057	0.000	0.000	-0.000	-0.000	-0.000
Anhydride	0.000	0.000	-0.022	-0.021	-0.022	0.098	0.343
Epoxide	-0.000	-0.000	-0.007	0.018	-0.012	0.107	-0.043
Thioester	-0.016	-0.029	0.000	-0.000	0.000	0.000	-0.000
Oxime	-0.010	-0.005	0.026	0.022	-0.063	0.011	0.048
Oxazolidinone	-0.001	0.025	-0.000	0.000	0.000	0.000	-0.000
Urea	-0.031	0.050	-0.010	0.201	-0.042	-0.028	0.065
Guanidine	-0.281	-0.014	0.000	-0.000	0.000	0.000	-0.000
Ether	0.069	0.399	-0.000	0.000	0.000	0.000	0.000
Sulfonamide	-0.005	-0.057	0.000	-0.000	-0.000	-0.000	-0.000
Sulfone	-0.009	-0.183	0.000	-0.000	-0.000	-0.000	-0.000
N-Oxide	-0.021	-0.018	0.703	-0.003	0.017	0.021	-0.028
Nitrile	0.000	0.000	0.035	0.027	-0.036	0.333	0.843
Thiol	-0.000	0.000	-0.008	0.051	-0.004	-0.194	-0.131
Thioether	-0.103	-0.051	0.010	0.577	-0.056	-0.053	-0.008
Fluorine	0.010	0.080	-0.000	0.000	0.000	0.000	0.000
Pyridine	0.024	0.046	-0.000	0.000	-0.000	-0.000	0.000
Alkyl halide	0.008	0.027	-0.000	-0.000	0.000	0.000	-0.000
Aryl halide	-0.067	0.002	0.000	0.000	0.000	-0.000	0.000
Alkene	-0.111	0.030	-0.000	-0.000	0.000	-0.000	0.000
Alkylgreater than5 C	0.211	-0.016	0.000	0.000	-0.000	-0.000	-0.000
Phosphonate	-0.123	0.055	0.006	-0.127	0.027	0.018	-0.041
Hydrozone	0.140	-0.047	-0.008	-0.478	0.046	0.044	0.007
Other_1	-0.026	-0.004	0.027	0.165	0.159	-0.639	0.271
Phosphate	-0.026	-0.004	-0.025	0.266	-0.211	0.595	-0.246
Carbamate	-0.015	0.051	-0.000	0.000	0.000	-0.000	0.000
Nitro	-0.284	-0.017	-0.008	-0.478	0.046	0.044	0.007
Nitrate	-0.021	-0.018	-0.707	0.003	-0.017	-0.021	0.028
Steroid	0.016	-0.146	0.000	0.000	-0.000	-0.000	0.000
Hormone	0.043	-0.022	0.000	-0.000	-0.000	0.000	-0.000
O-heterocyclic	0.056	0.037	-0.000	0.000	0.000	-0.000	0.000
N-heterocyclic	-0.083	-0.123	0.000	-0.000	-0.000	-0.000	0.000
S-heterocyclic	0.046	0.044	-0.000	0.000	0.000	-0.000	0.000
Long alkyl	0.061	-0.005	0.000	0.000	0.000	0.000	-0.000
Phenyl ring	0.047	0.011	0.022	-0.096	-0.674	-0.172	0.034
Erythromycin deriv	-0.010	-0.005	-0.025	-0.021	0.062	-0.010	-0.047
Tetracycline	0.310	0.005	0.000	0.000	-0.000	0.000	-0.000

Macrocyclic	-0.278	0.052	-0.000	-0.000	0.000	0.000	0.000
Macrolide	-0.102	-0.454	0.000	-0.000	-0.000	-0.000	-0.000
Benzodiazepine	-0.009	-0.058	0.000	0.000	-0.000	-0.000	-0.000
Barbiturate	0.039	0.015	0.009	-0.179	0.038	0.025	-0.058
Water	0.003	0.001	0.000	0.000	0.000	0.000	0.000
Ethanol	-0.027	-0.062	0.000	-0.000	-0.000	-0.000	-0.000
HCl	0.009	0.027	-0.000	0.000	0.000	0.000	0.000
Na+	0.026	0.000	-0.000	0.000	-0.000	-0.000	0.000
Gd3+	0.060	-0.503	0.000	0.000	-0.000	-0.000	-0.000

Variable	PC57	PC58	PC59	PC60	PC61	PC62	PC63
Primary	0.009	0.016	-0.000	-0.026	-0.000	0.000	-0.065
Secondary	0.000	-0.000	0.000	-0.000	-0.000	0.000	-0.000
Tertiary	-0.000	-0.000	0.000	0.000	0.000	-0.000	-0.000
Aromatic/enamine	0.000	-0.000	-0.000	0.000	0.000	-0.000	0.000
Primary_1	0.000	-0.000	0.000	-0.000	0.000	0.000	-0.000
Secondary_1	-0.000	-0.000	0.000	0.000	0.000	-0.000	-0.000
Tertiary_1	-0.000	-0.000	0.000	0.000	0.000	-0.000	-0.000
Vinyl alcohol	0.000	0.000	-0.000	0.000	0.000	-0.000	0.000
Phenol	-0.000	0.000	-0.000	0.000	0.000	-0.000	0.000
Carboxylic	0.000	0.000	-0.000	-0.000	-0.000	0.000	-0.000
Sulfonated	-0.000	-0.000	0.000	-0.000	-0.985	-0.170	0.000
Other	0.000	0.000	0.000	-0.000	0.170	-0.985	-0.000
Ketone	-0.000	-0.000	0.000	0.000	0.000	-0.000	-0.000
Aldehyde	0.053	0.148	0.975	0.154	0.000	0.000	0.000
Enone	-0.000	0.000	0.000	-0.000	0.000	0.000	-0.000
Ester	0.000	0.000	0.000	-0.000	-0.000	0.000	-0.000
1o amide	-0.000	0.000	-0.000	0.000	0.000	-0.000	-0.000
2o amide	-0.000	0.000	0.000	-0.000	0.000	0.000	-0.000
3o amide	-0.000	-0.000	-0.000	0.000	0.000	-0.000	0.000
Anhydride	-0.371	-0.166	0.185	-0.817	0.000	0.000	0.061
Epoxide	0.402	-0.900	0.109	0.022	0.000	-0.000	0.038
Thioester	0.000	0.000	-0.000	-0.000	-0.000	0.000	0.000
Oxime	-0.003	-0.028	0.010	-0.022	0.000	0.000	-0.659
Oxazolidinone	-0.000	-0.000	0.000	-0.000	0.000	-0.000	-0.000
Urea	-0.020	-0.020	0.001	0.029	0.000	-0.000	-0.210
Guanidine	0.000	0.000	0.000	-0.000	-0.000	0.000	0.000
Ether	-0.000	-0.000	0.000	0.000	0.000	-0.000	-0.000
Sulfonamide	-0.000	0.000	-0.000	0.000	0.000	-0.000	0.000
Sulfone	0.000	0.000	-0.000	-0.000	-0.000	-0.000	0.000
N-Oxide	-0.005	-0.004	0.004	-0.023	-0.000	0.000	0.016
Nitrile	0.312	0.142	-0.046	0.217	0.000	-0.000	0.078
Thiol	0.771	0.313	-0.014	-0.497	-0.000	0.000	-0.026
Thioether	-0.037	-0.006	0.002	-0.001	0.000	0.000	0.103
Fluorine	-0.000	-0.000	0.000	-0.000	0.000	0.000	-0.000
Pyridine	-0.000	-0.000	-0.000	0.000	0.000	-0.000	-0.000
Alkyl halide	0.000	0.000	0.000	-0.000	-0.000	0.000	-0.000
Aryl halide	-0.000	-0.000	0.000	0.000	0.000	-0.000	-0.000
Alkene	-0.000	0.000	0.000	-0.000	0.000	0.000	-0.000
Alkylgreater than5 C	0.000	-0.000	0.000	-0.000	-0.000	0.000	0.000
Phosphonate	0.013	0.013	-0.001	-0.018	-0.000	0.000	0.133
Hydrozone	0.031	0.005	-0.002	0.000	-0.000	-0.000	-0.085

Other_1	-0.039	-0.101	0.023	0.071	0.000	-0.000	-0.012
Phosphate	0.008	0.089	-0.021	-0.059	-0.000	0.000	-0.022
Carbamate	-0.000	-0.000	0.000	0.000	0.000	-0.000	-0.000
Nitro	0.031	0.005	-0.002	0.000	-0.000	-0.000	-0.085
Nitrate	0.006	0.004	-0.004	0.023	0.000	-0.000	-0.016
Steroid	0.000	-0.000	0.000	0.000	0.000	-0.000	0.000
Hormone	0.000	0.000	-0.000	-0.000	-0.000	0.000	0.000
O-heterocyclic	0.000	-0.000	0.000	0.000	0.000	-0.000	-0.000
N-heterocyclic	-0.000	-0.000	0.000	-0.000	0.000	0.000	0.000
S-heterocyclic	0.000	-0.000	0.000	0.000	0.000	-0.000	-0.000
Long alkyl	0.000	-0.000	0.000	-0.000	-0.000	-0.000	0.000
Phenyl ring	-0.009	-0.016	0.000	0.026	0.000	-0.000	0.065
Erythromycin deriv	0.003	0.028	-0.009	0.022	-0.000	-0.000	0.650
Tetracycline	0.000	-0.000	0.000	-0.000	-0.000	0.000	-0.000
Macrocyclic	-0.000	0.000	0.000	-0.000	0.000	0.000	-0.000
Macrolide	0.000	0.000	-0.000	-0.000	-0.000	0.000	0.000
Benzodiazepine	-0.000	0.000	0.000	-0.000	-0.000	0.000	0.000
Barbiturate	0.018	0.018	-0.001	-0.026	-0.000	0.000	0.188
Water	-0.000	0.000	-0.000	0.000	0.000	-0.000	0.000
Ethanol	-0.000	0.000	0.000	-0.000	-0.000	0.000	0.000
HCl	0.000	-0.000	-0.000	0.000	-0.000	-0.000	-0.000
Na+	-0.000	-0.000	0.000	0.000	0.000	-0.000	-0.000
Gd3+	-0.000	0.000	-0.000	0.000	0.000	0.000	0.000

Variable	PC64
Primary	0.006
Secondary	0.000
Tertiary	0.000
Aromatic/enamine	0.000
Primary_1	-0.000
Secondary_1	0.000
Tertiary_1	0.000
Vinyl alcohol	0.000
Phenol	-0.000
Carboxylic	0.000
Sulfonated	0.000
Other	-0.000
Ketone	0.000
Aldehyde	0.001
Enone	0.000
Ester	0.000
1o amide	0.000
2o amide	0.000
3o amide	0.000
Anhydride	0.014
Epoxide	-0.002
Thioester	-0.000
Oxime	-0.252
Oxazolidinone	0.000
Urea	0.586

Guanidine	-0.000
Ether	0.000
Sulfonamide	-0.000
Sulfone	-0.000
N-Oxide	0.025
Nitrile	-0.056
Thiol	0.040
Thioether	-0.207
Fluorine	0.000
Pyridine	0.000
Alkyl halide	0.000
Aryl halide	0.000
Alkene	-0.000
Alkylgreater than5 C	-0.000
Phosphonate	-0.370
Hydrozone	0.171
Other_1	0.039
Phosphate	0.103
Carbamate	0.000
Nitro	0.171
Nitrate	-0.025
Steroid	-0.000
Hormone	-0.000
O-heterocyclic	0.000
N-heterocyclic	-0.000
S-heterocyclic	0.000
Long alkyl	0.000
Phenyl ring	-0.006
Erythromycin deriv	0.249
Tetracycline	-0.000
Macrocyclic	-0.000
Macrolide	-0.000
Benzodiazepine	0.000
Barbiturate	-0.524
Water	0.000
Ethanol	-0.000
HCl	0.000
Na+	0.000
Gd3+	-0.000

Figure I Principle components for Chemical functional groups and structural properties in table form.

Table I

Variable of interest	Principal components identified as showing variability
Primary amine group	C1, C2, C4 (C7)
Secondary amine group	C1, C3, C6
Aromatic/enamine group	C1, C2, C6
Tertiary alcohol (OH structure)	C1, C2, C3, C4
Vinyl alcohol group	C2, C3, C5
Carboxylic acid group	C3, C4, C5, C6
Secondary amide (carbonyl group)	C1, C2, C4 (C7)
Oxime (N group)	C2, C3, C4, (C12, C14)
Phosphonate group	C1, C2
Hydrozone group	C1, C2, C3
Phosphate	C1 C2
Phenyl ring	C1, C2, C3, C4 (C7)
Macrolide	C2, C3, C4
Na+ Associated group	C1, C2
Gd3+ Associated group	C3, C4, C5, C6

Table I The data shown in brackets indicates variability within the data for the identified variable post principal component 6. Principal components account for 38% of the variability seen in the dataset.

Table II

Structural feature or Functional Group Identified	Pharmaceutical Product Identified
Primary amine	Aluvial, Levothyroxine, Gabapentin, Cycloserine, Sevelamer, Folic acid
Secondary amine	Tamsulosin, Sumatriptan base, Sevelamer, Salmeterol xinafoate, Oxis, Metronazole, Marcaine, Gopten, Gadopentetate monomeglumine, Gadopentetate dimeglumine, Furosemide, Bambec
Aromatic /enamine	Folic acid, HPMPC, Hytrin, Deflox, Lupron, Nizatidine, Plendil, Ranitidine
Tertiary alcohol (OH structure)	Betamethasone acetate, Betamethasone disodium phosphate, Calcijex, Clarithromycin, Doxycycline hyclate, Doxycycline monohydrate, Invermectin, Klacid, Paricalcitol, Roxithromycin, Sevelamar
Vinyl alcohol	Doxycycline hyclate, Doxycycline monohydrate, Warfarin
Carboxylic acid	Blopress, Brofen, Epival, Folic acid, Furosemide, Gadopentetate dimeglumie, Gadopentetate monomeglumie, Gopten, Ketoprofen, Levothyroxine, Quinapril, Salmeterol xinafoate, Teveten
Secondary amide (Carbonyl)	Ciclosporin, Citanest, Folic acid, Iodixanol, Iohexol, Iopamidol, Lupron, Marcaine, Oxis, Quinapril

Structural feature or Functional Group Identified	Pharmaceutical Product Identified
Oxime (N group)	Roxithromycin
Phosphonate	Betamethasone disodium phosphate, HPMPC
Hydrozone	Betamethasone disodium phosphate, Teveten
Phosphate	Betamethasone disodium phosphate
Phenyl ring	Atenolol, Bambec, Blopress, Brofen, Citanest, Deflox, Furosemide, Gopten, Hytrin, Levothyroxine, Lupron, Marcaine, Meperidine, Metronazole, Oxis, Plendil, Warfarin
Macrolide	Clarithromycin, Invermectin, Roxithromycin, Klacid
Na+ Association	Betamethasone disodium phosphate, Epival
Gd3+ Association	Gadopentetate monomeglumine, Gadopentetate dimeglumine

Table II Identified chemical functional groups and structural properties in relation to the pharmaceutical products used for PCA.

Table III

Pharmaceutical Products not represented by the features identified as giving high variation amongst the dataset and their functional and structural features	
Pharmaceutical product	Structural and functional group information
Advicor	Ketone, Ether, Alkyl >5 carbons, N- heterocyclic

Pharmaceutical Products not represented by the features identified as giving high variation amongst the dataset and their functional and structural features	
Pharmaceutical product	Structural and functional group information
Androgel	Secondary alcohol, Enone, Steroid
Beclomethasone dipropionate	Secondary alcohol, Ketone, Ester, Steroid, Alkyl halide
Beclomethasone dipropionate monohydrate	Secondary alcohol, Water, Steroid, Alkyl halide, Ester, Ketone
Ciclesonide	Secondary alcohol, Ketone, Ester, Ether, Steroid
Clobetasol proprionate	Secondary alcohol, Ketone, Ester, Fluorine, Steroid, Alkyl halide
Conholip	Ketone, Ester, Alkyl >5 carbons
Dexamethasone dipropionate	Secondary alcohol, Ketone, Fluorine, Ester, Ether, Steroid
Fluticasone furoate	Secondary alcohol, Ketone, Fluorine, Ester, Thioester, Ether, Steroid
Fluticasone propionate	Secondary alcohol, ketone, Ester, Thioester, Steroid
Halobetasol	Steroid, Alkyl halide, Fluorine, Ester, Ketone, Secondary alcohol
Imdur	Secondary alcohol, Ether, Nitrate, O-heterocyclic
Isoflurane	Ether

Pharmaceutical Products not represented by the features identified as giving high variation amongst the dataset and their functional and structural features	
Pharmaceutical product	Structural and functional group information
Isradipine	Ester, Pyridine, Alkyl >5 carbons, N-heterocyclic
Meprobamate	Carbamate
Methohexital	Urea, N-heterocyclic, Barbiturate
Mometasone furoate anhydrous	Secondary alcohol, Ketone, Ester, Ether, Alkyl halide, Steroid
Mometasone furoate monohydrate	Secondary alcohol, Ketone, Ester, Ether, Alkyl halide, Steroid, Water
Nimbex	Ester, Ether, Sulfone, N-heterocyclic
Olanzapine	Tertiary amine, Thioester
Progesterone	Ketone, Enone, Steroid, Hormone
Severane	Ether
Venlafaxin	Tertiary amine, Ether, Ethanol

Table III Pharmaceutical products not containing features identified as contributing to the data variation in the first six principle components.

Table IV

Identified group or prominent feature	Identified pharmaceutical products	Identified chemical or structural feature
1	Betamethasone disodium phosphate (9)	Na+ Association, Hydrozone, Phosphate, Phosphonate, Tertiary alcohol association, Secondary alcohol, Ketone, Aryl halide, Steroid
2	Clarithromycin (15)	Macrolide, Tertiary alcohol structure, tertiary amine, secondary alcohol, ketone, ester, ether
	Invermectin (42)	Macrolide, Tertiary alcohol structure, Secondary alcohol, Ester, Ether
	Doxcycline monohydrate (22)	Tertiary alcohol structure, Vinyl alcohol, Tertiary amine, Secondary alcohol, Ketone, Primary amide, Tetracycline
	Klacid (44)	Tertiary alcohol structure, Macrolide, Secondary alcohol, Tertiary amine, Ketone, Ester, Ether
3	Lupron (46)	Aromatic enamine, Secondary amide, Phenyl ring, Primary alcohol, Phenol, Secondary amide, Guanidine, Alkyl >5

		carbons, N-heterocyclic,
4	Doxcycline hyclate (21) Roxithromycin(63)	Tertiary alcohol structure, Vinyl alcohol, Tertiary amine, Tertiary alcohol, Ketone, Primary amide, Tetracycline Tertiary alcohol structure, Oxime group, Macrolide, Tertiary amine, Secondary alcohol, Ester, Oxime, Ether, Erythromycin derivative
5	Nizatidine (55)	Aromatic enamine, Tertiary amine, Thioester, Nitro, N-heterocyclic, S-heterocyclic
6	Levothyroxine (45)	Primary amine, Carboxyl acid, Phenyl ring, Phenol, Ether, Aryl halide, Hormone
7	Gadopentetate dimeglumine (29) Gadopentetate monomeglumine (30) Imdur (36)	Secondary amine, Carboxylic acid, Gd3+ association, Secondary amide, Primary alcohol, Secondary alcohol Secondary amine, Carboxylic acid, Gd3+ association, Secondary amide, Water, Tertiary amine, Primary alcohol, Secondary alcohol No significant group identified by scree plot analysis. Contains Secondary alcohol, Ether, Nitrate, O-heterocyclic features.
8	HPMPC (34) Teveten (69)	Phosphonate, Aromatic enamine, Primary alcohol, Urea Carboxylic acid, Hydrozone, Tertiary amine, thioester, thioether, N-heterocyclic
9	Epival (23) Isradipine (41)	Carboxylic acid, Na+ associated No significant group identified by scree plot analysis. Contains Ester, Pyridine, Alkyl >5 carbons, N-heterocyclic

10	Advicor (1)	No significant group identified by scree plot analysis. Contains Ether, Ketone, Alkyl >5 carbons, N-heterocyclic
	Androgel (3)	No significant group identified by scree plot analysis. Contains Secondary alcohol, Enome, Steroid
	Ciclesonide (13)	No significant group identified by scree plot analysis. Contains Secondary alcohol, Ketone, Ester, Ether, Steroid
	Conholip (17)	No significant group identified by scree plot analysis. Contains Ketone, Ester, Alkyl >5 carbons
	Progesterone (60)	No significant group identified by scree plot analysis. Contains Ketone, Enone, Steroid, Hormone
11	Beclomethasone dipropionate (6)	No significant group identified by scree plot analysis. Contains Secondary alcohol, Ketone, Ester, Steroid, Alkyl halide
	Beclomethasone dipropionate monohydrate (7)	No significant group identified by scree plot analysis. Contains Secondary alcohol, Ketone, Ester, Steroid, Alkyl halide, water
	Betamethasone acetate (8)	Tertiary alcohol, Secondary alcohol, Tertiary alcohol, Ketone, Ester, Fluorine, Steroid
	Clobetasol propionate (16)	No significant group identified by scree plot analysis. Contains Secondary alcohol, Ketone, Ester, Fluorine, Steroid, Alkyl halide
	Dexamethasone	No significant group identified by scree plot analysis. Contains Secondary alcohol,

11	dipropionate (20)	Ketone, Ester, Fluorine, Steroid features
	Fluticasone furoate (24)	No significant group identified by scree plot analysis. Contains Secondary alcohol, Ketone, Fluorine, Ester, Thioester, Ether, Steroid
	Fluticasone propionate (25)	No significant group identified by scree plot analysis. Contains Secondary alcohol, Ketone, Ester, Thioester, Steroid
	Halobetasol (33)	No significant group identified by scree plot analysis. Contains Steroid, Alkyl halide, Fluorine, Ester, Ketone, Secondary alcohol
	Mometasone furoate anhydrous (52)	No significant group identified by scree plot analysis. Contains Secondary alcohol, Ketone, Ester, Ether, Alkyl halide, Steroid
	Mometasone furoate monohydrate (53)	No significant group identified by scree plot analysis. Contains Secondary alcohol, Ketone, Ester, Ether, Alkyl halide, Steroid, Water
12	Aluvia (2)	Primary amine and also Ester, Primary amide, Alkyl >5 carbon (nine present)
	Nimbex (54)	No significant group identified by scree plot analysis. Contains Ester, Ether, Sulfone, N-heterocyclic
	Venlafaxine (70)	No significant group identified by scree plot analysis. Contains Tertiary amine, Ether, Ethanol
Main data set	Atenolol (4)	Phenyl Ring, Secondary amine, Secondary alcohol, Primary amide, Ether, Phenyl ring

	Bambec (5)	Phenyl Ring, Secondary amine, Secondary alcohol, Carbamate
	Blopress (10)	Carboxylic acid, Phenyl ring, Ether, N-heterocyclic
	Brofen (11)	Carboxylic acid, Phenyl ring,
	Calcijex (12)	Tertiary alcohol, Secondary alcohol, , Alkenes, Alkyl >5 carbons
	Citanest (14)	Phenyl ring, Secondary amine, Secondary amide
	Cycloserine (18)	Primary amine, Ketone, Oxazolidonone,
	Deflox (19)	Phenyl ring, Aromatic enamine, Tertiary amide, Guanidine, Ether, N-heterocyclic, Phenyl ring, Water
	Folic acid (26)	Carboxylic acid, Primary amine ,Secondary amide, , Secondary amide, N-heterocyclic
	Furosemide (27)	Secondary amine, Carboxylic acid, O-heterocyclic, Aryl halide, Sulfonamide,
	Gabapentin (28)	Primary amine, Secondary amide, Ester
	Ciclosporin (31)	Secondary amide, Secondary alcohol, Secondary amide, Tertiary amide, Alkyl >5C, Macrocyclic
		Phenyl ring, Aromatic enamine, Tertiary

Main dataset	Hytrin (35)	amide, Guanidine, Ether, N-heterocyclic
	Iodixanol (37)	Secondary amide, Primary alcohol, Secondary alcohol, Tertiary amide
	Iopanidol (39)	Secondary amide, Primary alcohol, Secondary alcohol
	Isoflurane (40)	No significant group identified by scree plot analysis. Contains Ether
	Marcaine (47)	Secondary amine, phenyl ring, secondary amide, O-heterocyclic
	Meperidine (48)	Phenyl ring, Tertiary amine, Ester, N-heterocyclic,
	Meprobamate (49)	No significant group identified by scree plot analysis. Contains Carbamate
	Methohexital (50)	No significant group identified by scree plot analysis. Contains Urea, N-heterocyclic, Barbituate
	Olanzapine (56)	No significant group identified by scree plot analysis. Contains Tertiary amine, Thioester
	Oxis (57)	Secondary amine, Secondary amide, Phenyl ring, Secondary alcohol, Phenol, Ether
		Tertiary alcohol, Secondary alcohol, Alkenes, Long alkyl

Main Dataset	Paricalcitol (58)	Phenyl ring, Aromatic enamine, Ester, Aryl halide, alkenes, N-heterocyclic
	Plendil (59)	Carboxylic acid, Secondary amide, Ester, Tertiary amide, Secondary amide, Ether, N-heterocyclic
	Quinapril (61)	Aromatic enamine, Tertiary amine, Thioether, Nitro, O-heterocyclic
	Ranitidine (62)	Secondary amine, Carboxylic acid, Phenol, Primary alcohol, Secondary alcohol, Ether, Long alkyl
	Salmeterol xinafoate (64)	No significant group identified by scree plot analysis. Contains Ether
	Severane (66)	Secondary amine, Ether, Sulfonamide
	Tamsulosin (68)	Vinyl alcohol, Phenyl ring, Ketone, Ester, O-heterocyclic
	Warfarin (71)	
Products not identified in score plot analysis	Sumatriptan Base (67)	Secondary amine, Tertiary amine, Sulfonamide, N-heterocyclic
	Sevelamer (65)	Primary amine, Secondary amine, Tertiary alcohol,
	Gopten (32)	Secondary amine, Carboxylic acid, Phenyl ring, Ester, Tertiary amide, N-heterocyclic

	Iohexol (38) Metronazole (51)	Secondary amide, Primary Alcohol, Secondary alcohol, Tertiary amide. Secondary amine, Phenyl ring, N-heterocyclic, Aryl halide, Sulfonamide, Tertiary amide

Table IV Identified pharmaceutical products and associated chemical or structural features identified by examination of the score plot and the significant features of the pharmaceutical products. (Where blue writing indicates the information is not an identified feature by scree plot analysis).

Figure II

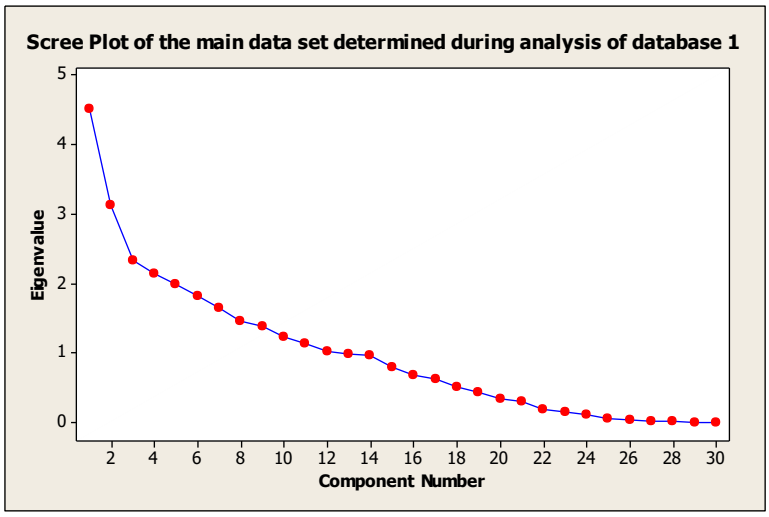


Figure II Scree plot of the main data set determined during initial PCA analysis of database 1.

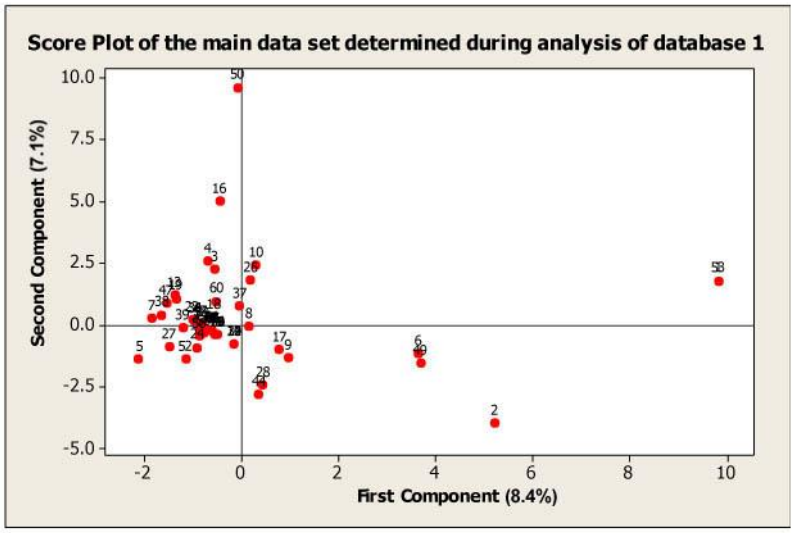


Figure III Score plot of the main data set determined during initial PCA analysis of database 1.

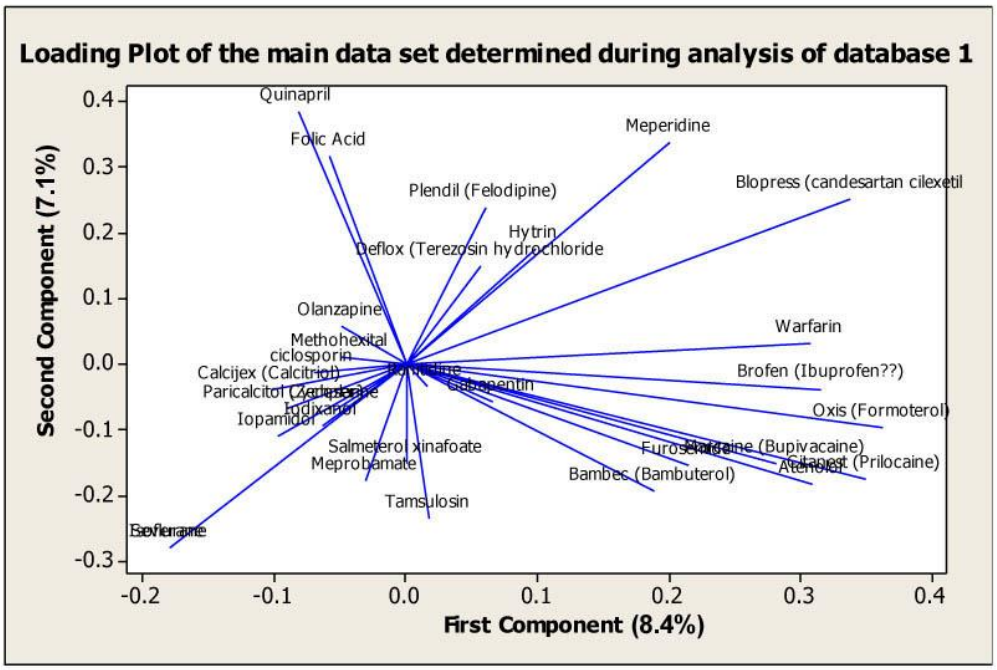


Figure IV Loading Plot of of the main data set determined during initial PCA analysis of database 1.

Table V

Cluster identified	Variables contained in cluster	API associated with the variables in the data set

Cluster identified	Variables contained in cluster	API associated with the variables in the data set
1	Tertiary alcohol group, Oxime group, macrolide, ether group	Betamethasone disodium phosphate, Betamethasone acetate, Clarithromycin, Paricalcitol, Calcijex, Doxycline hyclate, Doxycline monohydrate, klacid, Roxithromycin, Sevelmer, Roxithromycin, Clarithromycin, Klacid, Ivermectin, Roxithromycin, Mometasone furoate anhydrous, Mometasone furoate monohydrate, Nimbex, Oxis, Quinapril, Roxithromycin, Salmeterol xinafoate, Severane, Tamsulosin, Venlafaxine.
2	Primary amine group, alkyl halide group, guanidine group, phenol group	Aluvia, cycloserine, Folic acid, Gabapentine, Levothyroxine, sevelamer, Beclomethasone dipropionate, Beclomethasone dipropionate monohydrate, Clobetasol propionate, Halobetasol, Mometasone furoate anhydrous, Mometasone furoate monohydrate, deflox, hytrin, lupron , Levothyroxine, lupron, oxis
3	Phosphate group, Phosphonate group, Na+ group	Betamethasone disodium phosphate, Betamethasone disodium phosphate, HPMPC, Betamethasone disodium

Cluster identified	Variables contained in cluster	API associated with the variables in the data set
		phosphate, epival.
4	Tertiary amide group, aromatic /enamine group, N-heterocyclic group, alkenes group, pyridine group	Cyclosporin, Deflox, Gopten, Iodixanol, Iohexol, Metolazone, Quinapril, Deflox, Folic acid, HPMPC, Hytrin, Lupron, Nizatidine, Plendil, Ranitidine, Advicor, Blopress, Deflox, Eprosartan, Folic acid, Gopten, HPMPC, Hytrin, Isradipine, lupron, Meperidine, Methohexital, Metolazone, Nimbex, Nizatidine, Olanzapine, Plendil, Quinapril, Sumatriptan base, Calcijex, Paricalcitol, Plendil, Betamethasone disodium phosphate, Isradipine
5	Secondary amide group	Ciclosporin, Lupron, Citanest, Folic acid, Iodixanol, Iohexol, Marcaine, Iopamidol, Oxis, Quinapril
9	Carboxylic acid groups, Gd3+ structure, Thioether groups	Blopress, Epival, Eprosartan, Folic acid, Furosemide, Gadopentetate dimeglumine, Gadopentetate monomeglmine, Gopten, Ketoprofen, Levothyroxine, Quinapril, Salmeterol xinafoate, Gadopentetate dimeglumine, Gaopentetate monomeglmine,

Cluster identified	Variables contained in cluster	API associated with the variables in the data set
		Eprosartan, Nizatidine, Ranitidine
10	Erythromycin derivative structure, Ketone groups, Vinyl alcohol groups	Roxithromycin, Advicor, Beclomethasone dipropionate, Beclomethasone dipropionate monohydrate monohydrate, Betamethasone acetate, Betamethasone disodium phosphate, Ciclesonide, Clarithromycin, Clobetasol propionate, Conholip, Cycloserine, Dexamethosone dipropionate, Doxycycline monohydrate, Fluticasone furaroate, Fluticasone propionate, Halobetasol, ketoprofen, Klacid, Mometasone furoate anhydrous, Mometasone furoate monohydrate, Progesterone, Warfarin, Doxycycline hyclate, Doxycycline monohydrate, Warfarin
11	Steroid	Androgel, Beclomethasone dipropionate, Beclomethasone dipropionate monohydrate, Betamethasone acetate, Betamethasone disodium phosphate, Ciclesonide, Clobetasol propionate,

Cluster identified	Variables contained in cluster	API associated with the variables in the data set
		Dexamethosone dipropionate, Fluticasone furaroate, Fluticasone propionate, Halobetasol, Mometasone furoate anhydrous, Mometasone furoate monohydrate, Progesterone
12	Tertiary amine group	Clarithromycin, Doxycycline hyclate, Doxycycline monohydrate, Eprosartan, Gadopentetate dimeglumine, Gadopenetate monomeglumine, Klacid, Meperidine, Nizatidine, Olanzapine, Rantidine, Roxithromycin, Sumatriptan base, Venlafaxine
13	Thioester group	Eprosartan, Fluticasone furaroate, Fluticasone propionate, Olanzapine
14	Ester group	Aluvia, Beclomethasone dipropinate, Betamethasone dipropionate monohydrate, Betamethasone acetate, Ciclesonide, Clarithromycin, Clobetasol propionate, Conholip, Dexamethosone dipropionate, Fluticasone propionate, Fluticasone furaroate, Gabapentin, Gopten,

Cluster identified	Variables contained in cluster	API associated with the variables in the data set
		Halobetasol, Isradipine, Ivermectin, Klacid, Meperidine, Mometasone furoate anhydrous, Mometasone furoate monohydrate, Nimbex, Plendil, Quinapril, Roxithromycin, Warfarin.
15	Fluorine group	Betamethasone acetate, Betamethasone disodium phosphate, Clobetasol propionate, Dexamethosone dipropionate, Fluticasone furaroate, Fluticasone propionate, Halobetasol,
16	Enone group	Androgel, Progesterone,
17	Primary amide group	Aluvia, Atenolol, Doxyxycline hyclate, Doxycycline monohydrate,
18	Phenyl ring	Atenolol, Bambec, Blopress, Brofen, Citanest, Deflox, Furosemide , Gopten, hytrin, Marcaine, meparidine, metazone, oxis, plendil, warfarin, Levothyroxine, Marcaine
19	Hydrozone	Betamethasone disodium phosphate, Eprosartan

Table V Analysis of figure 5-7 score plot pc3 versus pc4 for Database 1 information. The API and the chemical functional groups and structural features it contains are shown in the same colour text.

Table VI

Cluster identified	Variables contained in cluster	API associated with the variables in the data set
1	Vinyl alcohol group	Doxycline hyclate, Doxycline monohydrate, Warfarin
2	Secondary amine group	Atenolol, Bambec, Citanest, Furosemide, Gabapentine monomeglumine, Gopten, Marcaine, Metolazone, Oxis, Salmeterol xinafoate, Sevelamer, Sumatriptan base, Tamsulosin,
3	Gd3+, Carboxylic acid group, Primary amide group	Gadopentetate dimeglumine, Gadopentetate monomeglumine, Blopress, Brofen, Epival, Eposartan, Folic acid, Flurosemide, Gadopentetate monomeglumine, Gopten, ketoprofen, Levthyroxine, Quinapril, Salmeterol xinafoate, Aluvia, Atenolol, Doxycycline hyclate, Doxcycline monohydrate.
4	Primary amine group, Hormone structural features, Phenol acid group, Sulfonamide group.	Aluvial, Atenolol, Doxycycline hyclate, Doxycycline monohydrate, Levothyroxine, Progesterone, Levothyroxine, Lupron, Oxis, Salmeterol xinafoate, Sumatriptan base, Tamsulosin, Metolazone, Doxycycline hyclate, Doxycycline monohydrate.

Cluster identified	Variables contained in cluster	API associated with the variables in the data set
5	Alkyl halide group	Beclomethasone dipropionate, Beclomethasone dipropionate monohydrate, Clobetasol propionate, Halobetasol, Mometasone furoate anhydrous, Mometasone furoate monohydrate.
6	Secondary amide, Ether group, Oxime group	Ciclosporin, Citanest, Folic acid, Fluticasone furaroate, Fluticasone propionate, Gabapentin, Gopten, Halobetasol, Isradipine, Klacid, Meperidine, Mometasone furoate anhydrous, Mometasone furoate monohydrate, Nimbex, Plendil, Quinapril, Roxithromycin, Warfarin. Advicor, Atenolol, Blopress, Ciclesonide, Clarithromycin, Deflox, Fluticasone furaroate, Hytrin, Imdur, Isoflurane, Ivermectin, Klacid, Levothyroxine, Mometasone furoate anhydrous, Mometasone furoate monohydrate, Nimbex, Oxis, Quinapril, Roxithromycin, Salmeterol xinafoate, Severane, Tamulosin, Venlafaxine. Roxithromycin
7	Ester group, Thioester	Aluvia, Beclomethasone dipropinate, Betamethasone dipropionate monohydrate, Betamethasone acetate, Ciclesonide, Clarithromycin, Clobetasol propionate, Conholip, Dexamethosone dipropionate, Fluticasone propionate, Fluticasone furaroate, Gabapentin, Gopten, Halobetasol, Isradipine, Ivermectin, Klacid, Meperidine, Mometasone furoate anhydrous, Mometasone furoate

Cluster identified	Variables contained in cluster	API associated with the variables in the data set
		monohydrate, Nimbex, Plendil, Quinapril, Roxithromycin, Warfarin., Fluticasone propionate, Fluticasone furaroate, Olazapine
8	N-heterocyclic structural features	Advicor, Blopress, Deflox, Eprosartan, Folic acid, Gopten, HPMPC, Hytrin, Isradipine, Lipron, Meperidine, Metolazone, Nimbex, Nizatidine, Olanzapine, Plendil, Quinapril, Sumatriptan base
9	Aromatic/enamine, Thioether, S-heterocyclic structural features, Nitro	Deflox, Folic acid, HPMPC, Hytrin, Lupron, Nizatidine, Plendil, Ranitidine, Eprosartan, Nizatidine, Ranitidine Nizatidine, Nizatidine
10	Guanidine	Deflox, Hytrin, Lupron,
11	Erythromycin derivative	Roxithromycin
12	Water group	Mometasone furoate monohydrate
13	HCL group, Tetracycline	Doxycycline hyclate, Doxycycline hyclate
14	Secondary amide group, Alkene group, Phenyl ring, Ethanol group	Ciclosporin, Citanest, Folic acid, Iodixanol, Iohexol, Iopamidol, Lupron, Marcaine, Oxis, Quinapril. Calcijex, Paricalcitol, Plendil. Warfarin, Plendil, Oxis, Methohexital, Mepridine, Marcaine, Levothyroxine, Lupron, Hytrin, Gopten, Furosemide, Deflox, Citanest, Brofen, Blopress, Bambec, Atenolol, Doxycycline hyclate, Venlafaxine

Cluster identified	Variables contained in cluster	API associated with the variables in the data set
15	Tertiary amine group, Primary alcohol OH group, Secondary alcohol OH group, Tertiary alcohol OH group, Ketone group, Enone group, Tertiary amide group. Oxazolidinone, Urea, Sulfone, N-oxide, Fluorine, Pyridine, Alkenes, Alkyl greater than 5 carbons, Phosphonate group, Hydrozone structure, Other functional groups, Phosphate, Carbamate, Steroid structural properties, O-heterocyclic structural properties, Long alkyl structures, Macro cyclic structures, Benzodiazepine structural features, Barbiturate structures, Na+ groups.	All remaining APIs analysed in the data were considered to be present in this group.

Table VI Variables associated with clusters identified in the analysis of figure 5-8.The API and the chemical functional groups and structural features it contains are shown in the same colour text.

Table VII

Cluster number	Features identified in the clusters on the Loading plot figure 5-11.
1	Phosphate, Na+, Hydrozone, Phosphonate Other
2	Aromatic enamine, Phenyl ring, Secondary amide, N-heterocyclic, Phenol, Guanidine
3	Secondary amine, Gd3+, Tertiary alcohol, Tertiary amine, Carboxylic acid also associated in this cluster but not distinctly
4	Macrolide, Ether, Erythromycin derivative, plus other secondary characteristics
5	Tertiary alcohol
6	Ketone and Steroid

Table VII Features identified in the clusters on the Loading plot (figure 5-10).

Table VIII

Variable number	**Variable of interest**	**Principle components showing variability**
1	Exact Mass	C1 C3
2	Molecular weight	C1 C3
3	C	C3 C4

Variable number	Variable of interest	Principle components showing variability
5	F	C2 C3
6	H	C3
8	N	C2 C4
9	Cl	C3
10	Boiling Point	C1
11	Melting Point	C1
12	Critical Temperature	C1
13	Critical Pressure	C1
14	Critical Volume	C1
15	Gibbs Energy	C3
16	Log P	C2 C4
17	MR	C1
18	Henry's Law	C2
19	Heat of Form	C3
20	tPSA	C1 C2 C3 C4
21	CLogP	C2
22	CMR	C1
23	ACD/LogP	C2

Variable number	Variable of interest	Principle components showing variability
24	ACD/LogD (pH5.5)	C2
25	ACD/BCF (pH5.5)	C2 C3 C4
26	ACD/KOC (pH5.5)	C2 C3 C4
27	H bond acceptors	C2 C3 C4
28	Freely rotating bonds	C4
29	Index of Refraction	C1
30	Molar volume	C1
31	Surface Tension	C1 C2
32	Flash Point	C1
33	Boiling Point	C1
34	ACD/LogD (pH7.4)	C2
35	ACD/BCF (pH7.4)	C2 C3 C4
36	ACD/KOC (pH7.4)	C2
37	H bond donors	C2
38	Polar surface area	C1 C2 C3 C4
39	Molar refractivity	C1
40	Polarizability	C1
41	Density	C4

Variable number	Variable of interest	Principle components showing variability
42	Enthalpy of vaporisation	C1
43	Vapour pressure	C3 C4

Table VIII Analysis of the first 4 Principal Components indicating which variabiles contribute to the variability. The Table can be interpreted as follows; the variable number refers to the number given to a specific variable this is named alongside it. The variables were identified by data analysis to add significantly to the variance. The third column identifies the principle component, which scored highly for the given variable.

Table IX

Identified Group or prominent feature	Identified Pharmaceutical products
1	Atenolol (3), Meprobamate (35), Gabapentin (20)
2	Meperidine (34), Brofen (7)
3	Isoflurane (28), Severane (49)
4	Progesterone (45)
5	Plendil (44)
6	Calcijex (8) Paricalcitol (43)
7	Ciclesonide (9)

Identified Group or prominent feature	Identified Pharmaceutical products
8	Fluticasone propionate (17)
9	Gopten (22), Quinapril (46), Halobetasol (23), Mometasone furoate monohydrate (38), Clobetasol propionate (12), Dexamethasone dipropionate (15)
10	Hytrin (25)
11	Warfarin (54), Marcaine (33), Androgel (2)
Products not identified in analysis of the score plot for principal components one and two	Aluvia (1), Beclomethasone dipropionate monohydrate (4), Betamethasone acetate (5), Blopress (6), Citanest (10), Clarithromycin (11), Cycloserine (13), Deflox (14), Fluticasone propionate (17), Folic acid (18), Furosemide (19), Ciclosporin (21), HPMPC (24), Imdur (26), Iodixanol (27), Ketoprofen (29), Klacid (30), Levothyroxine (31), Lupron (32), Methohexital (36), Metronazole (37), Nimbex (39), Nizatidine (40), Olanzapine (41), Oxis (42), Ranitidine (47), Salmeterol xinafoate (48), Sumatriptan base (50), Tamulosin (51), Teveten (52), Venafaxine (53)

Table IX Identified clusters and prominent features within score plot (figure 5-12) produced during PC analysis of Database 2.

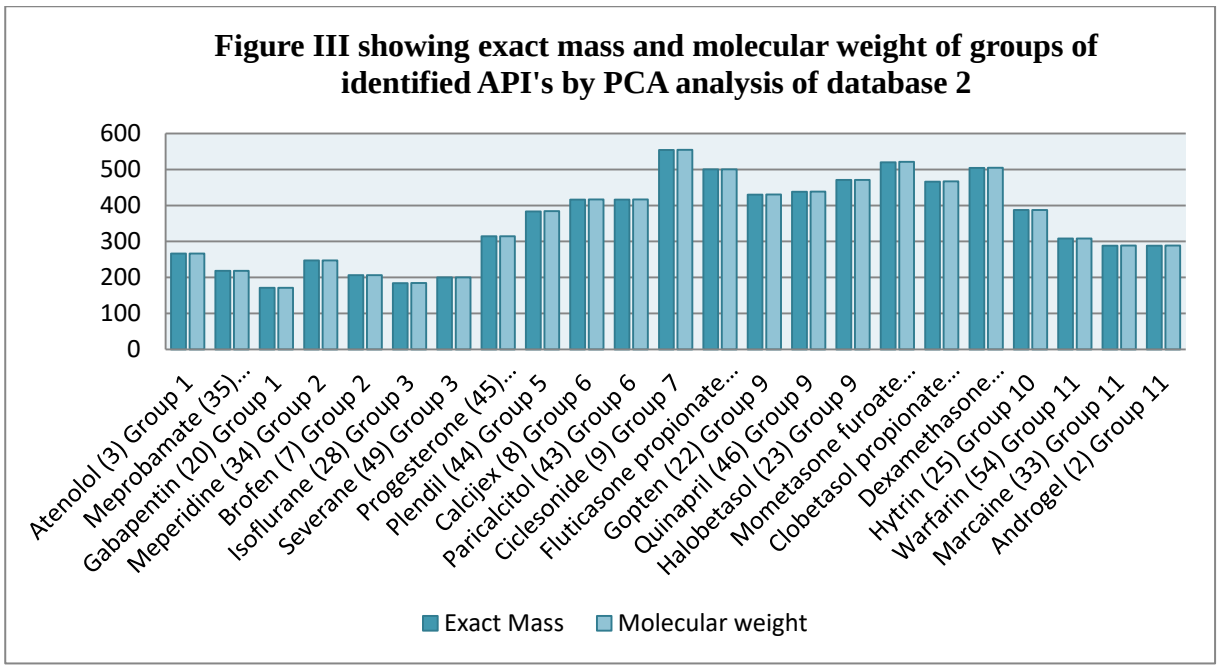


Figure III Exact mass and Molecular weight of products used in a PCA of Database 2. The figure shows products from a score plot of principal component 1 and principal component 2.

Physicochemical profile of API's identified in table IX

Atenonol

Atenonol is an API which has physicochemical properties including an exact mass of 266.16 and a molecular weight of 266.34. There are no Chlorine, Sulphur or Fluorine elements present but there are 63.13 Carbon, 18.02 Oxygen, 8.33 Hydrogen and 10.52 Nitrogen. It has a boiling point (K) at 841.7, a melting point (K) 524.88, a critical temperature (K) 887.27 and a critical pressure of 24.46. Atenonol has a critical volume of 806.5. It has a Gibbs energy value of -50 and a Log P value of 0.22. Atenonol has a MR 74.62cm^3/mol. It has a Henry's law value of 16.25 and it has a Heat of form value of -427.56. Atenonol also has a tSPA value of 84.58, a C Log P value of -0.1086 and a CMR value of 7.4783. This API has an ACD/Log P of 0.335, an ACD/Log D (pH5.5) value -2.75, an ACD/BCF (pH5.5) value of 1, and an ACD/KOC (pH5.5) value of 1 and 5 H bond acceptors. There are 9 freely rotating bonds and it has an Index of Refraction of 1.54 and a Molar Volume of 236.659. The surface tension associated with Atenonol is 45.019; it has a flash point of 261.059 °c, and a boiling point of 508.049°c. Atenonol has an ACD/Log D (pH7.4) of -1.76, an ACD/BCF (pH7.4) of 1, an ACD/KOC (pH7.4) of 4 and it has the ability to donate 4 H bonds. The polar surface area of Atenonol is 84.58 the molar refractivity of Atenonol is 74.257, a polarizability of 29.438, a density of 1.125, an enthalpy of vaporisation of 81.95 and a vapour pressure of 7.69E-10.

Meprobamate

Meprobamate is an API which has physicochemical properties including an exact mass of 218.13 and a molecular weight of 218.2. There are no Chlorine, Sulphur or Fluorine elements present but there are 49.53 Carbon, 29.32 Oxygen, 8.31 Hydrogen and 12.84 Nitrogen. It has a boiling point (K) at 663.79, a melting point (K) 444.29, a critical temperature (K) 755.19 and a critical pressure of 26.6. Meprobamate has a critical volume of 638.5. It has a Gibbs energy value of -460.1 and a Log P value of 1.06. Meprobamate has a MR 53.73cm^3/mol. It has a Henry's law value of 11.32 and it has a Heat of form value of -804.82. Meprobamate also has a tSPA value of 104.64, a C Log P value of 0.915 and a CMR value of 5.4666. This API has an ACD/Log P of 0.7, an ACD/Log D (pH5.5) value 0.7, an ACD/BCF (pH5.5) value of 2, and an ACD/KOC (pH5.5) value of 57.25 and 6 H bond acceptors. There are 8 freely rotating bonds and it has an Index of Refraction of 1.479 and a Molar Volume of 191.464. The surface tension associated with Meprobamate is 43.902; it has a flash point of 229.739 °c, and a boiling point of 434.212°c. Meprobamate has an ACD/Log D (pH7.4) of 0.7, an ACD/BCF (pH7.4) of 2, an ACD/KOC (pH7.4) of 57.25 and it has the ability to donate 4 H bonds. The polar surface area of Meprobamate is 104.64, the molar refractivity of Meprobamate is 54.331 and it has a polarizability value of 21.538, a density of 1.14 and a value for enthalpy of vaporisation of 69.029 and a vapour pressure of 4.66E+01.

Gabapentin

Gabapentin is an API which has physicochemical properties including an exact mass of 171.13 and a molecular weight of 171.24. There are no Chlorine, Sulphur or Fluorine elements present but there are 63.13 Carbon, 18.69 Oxygen, 10.01 Hydrogen and 8.18 Nitrogen. It has a boiling point (K) at 643.35, a melting point (K) 465.83, a critical temperature (K) 784.09 and a critical pressure of 35.18. Gabapentin has a critical volume of 523.5. It has a Gibbs energy value of -233.6 and a Log P value of 0.88. Gabapentin has a MR 45.22cm^3/mol. It has a Henry's law value of 8.13 and it has a Heat of form value of -476.01. Gabapentin also has a tSPA value of 63.32, a C Log P value of -0.66 and a CMR value of 4.7317. This API has an ACD/Log P of 1.083, an ACD/Log D (pH5.5) value -1.47, an ACD/BCF (pH5.5) value of 1, and an ACD/KOC (pH5.5) value of 1 and 3 H bond acceptors. There are 4 freely rotating bonds and it has an Index of Refraction of 1.489 and a Molar Volume of 161.825. The surface tension associated with Gabapentin is 47.09, it has a flash point of 143.967 °c, and a boiling point of 314.438°c. Gabapentin has an ACD/Log D (pH7.4) of -1.42, an ACD/BCF (pH7.4) of 1, an ACD/KOC (pH7.4) of 1 and it has the ability to

donate 3 H bonds. The polar surface area of Gabapentin is 63.32, the molar refractivity of Gabapentin is 46.696; it has a polarizability value of 18.512, a density of 1.058 and a value for enthalpy of vaporisation of 61.095 and a vapour pressure of 2.94E-10.

Meperidine

Meperidine is an API which has physicochemical properties including an exact mass of 247.16 and a molecular weight of 247.33. There are no Fluorine, Sulphur or Chlorine elements present but there are 72.84 Carbon, 12.94 Oxygen, 8.56 Hydrogen and 5.66 Nitrogen. It has a boiling point (K) at 655.66, a melting point (K) 394.33, a critical temperature (K) 809.83 and a critical pressure of 22.7. Meperidine has a critical volume of 750.5. It has a Gibbs energy value of 19.3 and a Log P value of 2.64. Meperidine has a MR 71.89cm^3/mol. It has a Henry's law value of 6.4 and it has a Heat of form value of -290.43. Meperidine also has a tSPA value of 29.54, a C Log P value of 2.227 and a CMR value of 7.2429. This API has an ACD/Log P of 2.185, an ACD/Log D (pH5.5) value -0.08, an ACD/BCF (pH5.5) value of 1, and an ACD/KOC (pH5.5) value of 1.98 and 3 H bond acceptors. There are 4 freely rotating bonds and it has an Index of Refraction of 1.52 and a Molar Volume of 234.243. The surface tension associated with Meperidine is 38.337; it has a flash point of 111.636 °c, and a boiling point of 328.866°c. Meperidine has an ACD/Log D (pH7.4) of 1.62, an ACD/BCF (pH7.4) of 7.25, an ACD/KOC (pH7.4) of 98.88 and it has the ability to donate 0 H bonds. The polar surface area of Meperidine is 29.54, the molar refractivity of Meperidine is 71.266, and it has a polarizability value of 28.252, a density of 1.04.85 and a value for enthalpy of vaporisation of 57.13 and a vapour pressure of 8.43E-07.

Brofen

Brofen is an API which has physicochemical properties including an exact mass of 206.13 and a molecular weight of 206.28. There are no Fluorine, Sulphur, Nitrogen or Chlorine elements present but there are 75.69 Carbon, 51.51 Oxygen and 8.8 Hydrogen. It has a boiling point (K) at 673.33, a melting point (K) 405.31, a critical temperature (K) 789.46 and a critical pressure of 23.91. Brofen has a critical volume of 667.5. It has a Gibbs energy value of -187.43 and a Log P value of 3.75. Brofen has a MR 61.2cm^3/mol. It has a Henry's law value of 5.21 and it has a Heat of form value of -447.42. Brofen also has a tSPA value of 37.3, a C Log P value of 3.679 and a CMR value of 6.124. This API has an ACD/Log P of 3.502, an ACD/Log D (pH5.5) value 2.38, an ACD/BCF (pH5.5) value of 20.4, and an

ACD/KOC (pH5.5) value of 144.58 and 2 H bond acceptors. There are 4 freely rotating bonds and it has an Index of Refraction of 1.519 and a Molar Volume of 200.339. The surface tension associated with Brofen is 38.678; it has a flash point of 216.702 °c, and a boiling point of 319.643°c. Brofen has an ACD/Log D (pH7.4) of 0.58, an ACD/BCF (pH7.4) of 1, an ACD/KOC (pH7.4) of 2.3 and it has the ability to donate 1 H bonds. The polar surface area of Brofen is 37.3, the molar refractivity of Brofen is 60.776, and it has a polarizability value of 24.093, a density of 1.03 and a value for enthalpy of vaporisation of 59.252 and a vapour pressure of 1.86E-04.

Isoflurane

Isoflurane is an API which has physicochemical properties including an exact mass of 183.97 and a molecular weight of 184.49. There are no Sulphur or Nitrogen elements present but there are 19.22 Carbon, 8.67 Oxygen, 51.49 Fluorine, 19.22 Chlorine and 1.09 Hydrogen. It has a boiling point (K) at 320.33, a melting point (K) 150.59, a critical temperature (K) 443.8 and a critical pressure of 32.65. Isoflurane has a critical volume of 337.5. It has a Gibbs energy value of -1118.64 and a Log P value of 2.47. Isoflurane has a MR 23.96cm^3/mol. It has a Henry's law value of 0.0126 and it has a Heat of form value of -1253.07. Isoflurane also has a tSPA value of 9.23, a C Log P value of 1.764 and a CMR value of 2.2908. This API has an ACD/Log P of 2.118, an ACD/Log D (pH5.5) value 2.12, an ACD/BCF (pH5.5) value of 23.96, and an ACD/KOC (pH5.5) value of 338.11 and 1 H bond acceptors. There are 2 freely rotating bonds and it has an Index of Refraction of 1.301 and a Molar Volume of 123.843. The surface tension associated with Isoflurane is 15.828; it has a flash point of 10.643 °c, and a boiling point of 48.49°c. Isoflurane has an ACD/Log D (pH7.4) of 2.12, an ACD/BCF (pH7.4) of 23.96, an ACD/KOC (pH7.4) of 338.11 and it has the ability to donate 0 H bonds. The polar surface area of Isoflurane is 9.23. The molar refractivity of Isoflurane is 23.244; it also has a polarizability value of 9.215, a density of 1.49 and a value for enthalpy of vaporisation of 28.001 and a vapour pressure of 3.23E+02.

Severane

Severane is an API which has physicochemical properties including an exact mass of 200.01 and a molecular weight of 200.05. There are no Sulphur, Nitrogen or Chlorine elements present but there are 24.01 Carbon, 8 Oxygen, 66.48 Fluorine and 1.51 Hydrogen. It has a boiling point (K) at 301.53, a melting point (K) 150.54, a critical temperature (K) 383.12 and a critical pressure of 28.05. Severane has a critical volume of 375.5. It has a Gibbs energy value of -1482.63 and a Log P value of 2.24. Severane has a MR 23.78cm^3/mol. It has a

Henry's law value of -0.88 and it has a Heat of form value of -1653.66. Severane also has a tSPA value of 9.23, a C Log P value of 1.451 and a CMR value of 2.2942. This API has an ACD/Log P of 2.498, an ACD/Log D (pH5.5) value 2.5, an ACD/BCF (pH5.5) value of 46.59, and an ACD/KOC (pH5.5) value of 544.23 and 1 H bond acceptors. There are 2 freely rotating bonds and it has an Index of Refraction of 1.266 and a Molar Volume of 139.532. The surface tension associated with Severane is 13.027; it has a flash point of 11.446 °c, and a boiling point of 49.472°c. Severane has an ACD/Log D (pH7.4) of 2.5, an ACD/BCF (pH7.4) of 46.59, an ACD/KOC (pH7.4) of 544.23 and it has the ability to donate 0 H bonds. The polar surface area of Severane is 9.23. The molar refractivity of Severane is 23.362; it also has a polarizability value of 9.261, a density of 1.434 and a value for enthalpy of vaporisation of 28.084 and a vapour pressure of 3.11E+02.

Progesterone

Progesterone is an API which has physicochemical properties including an exact mass of 314.22 and a molecular weight of 314.46. There are no Fluorine, Sulphur, Nitrogen or Chlorine elements present but there are 80.21 Carbon, 10.18 Oxygen, and 9.62 Hydrogen. It has a boiling point (K) at 845.36, a melting point (K) 550.84, a critical temperature (K) 868.2 and a critical pressure of 16.71. Progesterone has a critical volume of 992.5. It has a Gibbs energy value of 50.86 and a Log P value of 3.78. Progesterone has a MR 92.44cm^3/mol. It has a Henry's law value of 5.58 and it has a Heat of form value of -430.54 also Progesterone has a tSPA value of 34.14, a C Log P value of 0.485839 and a CMR value of 9.3296. This API has an ACD/Log P of 3.827, an ACD/Log D (pH5.5) value 3.83, an ACD/BCF (pH5.5) value of 476.94, and an ACD/KOC (pH5.5) value of 2876.39 and 2 H bond acceptors. There are 1 freely rotating bonds and it has an Index of Refraction of 1.542 and a Molar Volume of 288.952. The surface tension associated with Progesterone is 41.171; it has a flash point of 166.683 °c, and a boiling point of 447.151°c. Progesterone has an ACD/Log D (pH7.4) of 3.83, an ACD/BCF (pH7.4) of 476.94, an ACD/KOC (pH7.4) of 2876.39 and it has the ability to donate 0 H bonds. The polar surface area of Progesterone is 34.14. The molar refractivity of Progesterone is 90.955; it also has a polarizability value of 36.057, a density of 1.088 and a value for enthalpy of vaporisation of 70.544 and a vapour pressure of 2.69E-06.

Plendil

Plendil is an API which has physicochemical properties including an exact mass of 383.25 and a molecular weight of 384.25. There are no Fluorine or Sulphur elements present but there are 56.26 Carbon, 16.66 Oxygen, 4.98 Hydrogen, 3.65 Nitrogen and 18.45 Chlorine. It

has a boiling point (K) at 925.72, a melting point (K) 652.09, a critical temperature (K) 907.19 and a critical pressure of 16.43. Plendil has a critical volume of 1027.5. It has a Gibbs energy value of -317.21 and a Log P value of 2.24. Plendil has a MR 98.5cm^3/mol. It has a Henry's law value of 1.35E-11 and it has a Heat of form value of -705.49 also Plendil has a tSPA value of 64.63, a C Log P value of 5.2968 and a CMR value of 9.9071. This API has an ACD/Log P of 4.761, an ACD/Log D (pH5.5) value 4.76, an ACD/BCF (pH5.5) value of 2440.56, and an ACD/KOC (pH5.5) value of 9250.27 and 5 H bond acceptors. There are 6 freely rotating bonds and it has an Index of Refraction of 1.55 and a Molar Volume of 300.844. The surface tension associated with Plendil is 42.194; it has a flash point of 238.964 °c, and a boiling point of 471.516°c. Plendil has an ACD/Log D (pH7.4) of 4.76, an ACD/BCF (pH7.4) of 2444.58, an ACD/KOC (pH7.4) of 9265.53 and it has the ability to donate 1 H bonds. The polar surface area of Plendil is 64.63. The molar refractivity of is 90.95595.782; it Plendil also has a polarizability value of 37.971, a density of 1.277 and a value for enthalpy of vaporisation of 73.428 and a vapour pressure of 0.

Calcijex

Calcijex is an API which has physicochemical properties including an exact mass of 416.33 and a molecular weight of 416.64. There are no Fluorine, Sulphur, Nitrogen or Chlorine elements present but there are 77.83 Carbon, 11.52 Oxygen, and 10.64 Hydrogen. It has a boiling point (K) at 1139.41, a melting point (K) 645.95, a critical temperature (K) 966.71 and a critical pressure of 11.8. Calcijex has a critical volume of 1356.5. It has a Gibbs energy value of -0.86 and a Log P value of 4.49. Calcijex has a 126.14MR cm^3/mol. It has a Henry's law value of 4.9 and it has a Heat of form value of -679.03 also Calcijex has a tSPA value of 60.69, a C Log P value of 4.475 and a CMR value of 12.8549. This API has an ACD/Log P of 5.632, an ACD/Log D (pH5.5) value 5.63, an ACD/BCF (pH5.5) value of 11219.63, and an ACD/KOC (pH5.5) value of 27578.06 and 3 H bond acceptors. There are 9 freely rotating bonds and it has an Index of Refraction value of 1.547 and a Molar Volume of 391.894. The surface tension associated with is 44.083; Calcijex has a flash point of 238.428 °c, and a boiling point of 565.009°c. Calcijex has an ACD/Log D (pH7.4) of 5.63, an ACD/BCF (pH7.4) of 11219.63, an ACD/KOC (pH7.4) of 27578.06 and it has the ability to donate 3 H bonds. The polar surface area of Calcijex is 60.69. The molar refractivity of Calcijex is 124.354; it also has a polarizability value of 49.298, a density of 1.063 and a value for enthalpy of vaporisation of 97.508 and a vapour pressure of 1.19E-12.

Paricalcitrol

Paricalcitrol is an API which has physicochemical properties including an exact mass of 416.33 and a molecular weight of 416.64. There are no Fluorine, Sulphur, Nitrogen or Chlorine elements present but there are 77.83 Carbon, 11.52 Oxygen, and 10.64 Hydrogen. It has a boiling point (K) at 1143.97, a melting point (K) 612.19, a critical temperature (K) 963.36 and a critical pressure of 12.14. Paricalcitrol has a critical volume of 1346.5. It has a Gibbs energy value of 23.84 and a Log P value of 4.52. Paricalcitrol has a 127.99MR cm^3/mol. It has a Henry's law value of 4.81 and it has a Heat of form value of -651.33 also Paricalcitrol has a tSPA value of 60.69, a C Log P value of 5.688 and a CMR value of 12.7029. This API has an ACD/Log P of 5.899, an ACD/Log D (pH5.5) value 5.9, an ACD/BCF (pH5.5) value of 17930.17, and an ACD/KOC (pH5.5) value of 38574.72 and 3 H bond acceptors. There are 8 freely rotating bonds and it has an Index of Refraction of 1.609 and a Molar Volume of 371.436. The surface tension associated with is 54.659; Paricalcitrol has a flash point of 238.344 °c, and a boiling point of 564.843°c. Paricalcitrol has an ACD/Log D (pH7.4) of 5.9, an ACD/BCF (pH7.4) of 17930.17, an ACD/KOC (pH7.4) of 38574.72 and it has the ability to donate 3 H bonds. The polar surface area of Paricalcitrol is 60.69. The molar refractivity of Paricalcitrol is 128.646; it also has a polarizability value of 50.999, a density of 1.122 and a value for enthalpy of vaporisation of 97.485 and a vapour pressure of 8.61E-14.

Ciclesonide

Ciclesonide is an API which has physicochemical properties including an exact mass of 554.32 and a molecular weight of 554.71. There are no Fluorine, Sulphur, Nitrogen or Chlorine elements present but there are 71.45 Carbon, 20.19 Oxygen, and 5.59 Hydrogen. It has a boiling point (K) at 1340.13, a melting point (K) 868.87, a critical temperature (K) 1098.37 and a critical pressure of 10.01. Ciclesonide has a critical volume of 1617.5. It has a Gibbs energy value of -375.72 and a Log P value of 3.97. Ciclesonide has a 152.9MR cm^3/mol. It has a Henry's law value of 16.55 and it has a Heat of form value of -1257.55 also Ciclesonide has a tSPA value of 99.13, a C Log P value of 5.87195 and a CMR value of 15.2391. This API has an ACD/Log P of 6.13, an ACD/Log D (pH5.5) value 6.13, an ACD/BCF (pH5.5) value of 26845.77, and an ACD/KOC (pH5.5) value of 51496.1 and 7 H bond acceptors. There are 7 freely rotating bonds and it has an Index of Refraction of 1.576 and a Molar Volume of 436.998. The surface tension associated with it is 51.861; Ciclesonide has a flash point of 209.975 °c, and a boiling point of 664.979°c. Ciclesonide has an

ACD/Log D (pH7.4) of 6.13, an ACD/BCF (pH7.4) of 26845.77, an ACD/KOC (pH7.4) of 51496.1 and it has the ability to donate 1 H bonds. The polar surface area of Ciclesonide is 99.13. The molar refractivity of Ciclesonide is 144.517; it also has a polarizability value of 57.291, a density of 1.237 and a value for enthalpy of vaporisation of 111.927 and a vapour pressure of 0.

Fluticasone propionate

Fluticasone propionate is an API which has physicochemical properties including an exact mass of 500.18 and a molecular weight of 500.57. There are no Nitrogen or Chlorine elements present but there are 59.99 Carbon, 15.9 Oxygen, 11.39 Fluorine, 6.24 Hydrogen and 6.41 Sulphur. It has a boiling point (K) at 1139.5, a melting point (K) 771.08, a critical temperature (K) 976.44 and a critical pressure of 12.54. Fluticasone propionate has a critical volume of 1348.5. It has a Gibbs energy value of -918.11 and a Log P value of 3.12. Fluticasone propionate has a 122.92MR cm^3/mol. It has a Henry's law value of 4.34 and it has a Heat of form value of -1501.83 also Fluticasone propionate has a tSPA value of 80.67, a C Log P value of 3.0326 and a CMR value of 12.5188. This API has an ACD/Log P of 3.73, an ACD/Log D (pH5.5) value 3.73, an ACD/BCF (pH5.5) value of 402.74, and an ACD/KOC (pH5.5) value of 258.47 and 5 H bond acceptors. There are 7 freely rotating bonds and it has an Index of Refraction of 1.556 and a Molar Volume of 377.027. The surface tension associated with it is 48.063; Fluticasone propionate has a flash point of 297.491 °c, and a boiling point of 568.289°c. Fluticasone propionate has an ACD/Log D (pH7.4) of 3.73, an ACD/BCF (pH7.4) of 402.73, an ACD/KOC (pH7.4) of 2548.45 and it has the ability to donate 1 H bonds. The polar surface area of Fluticasone propionate is 105.97. The molar refractivity of Fluticasone propionate is 121.148; it also has a no given polarizability value, a density of 1.328 and a value for enthalpy of vaporisation of 97.97 and a vapour pressure of 0.

Gopten

Gopten is an API which has physicochemical properties including an exact mass of 430.25 and a molecular weight of 430.54. There are no Fluorine, Sulphur or Chlorine elements present but there are 66.95 Carbon, 18.58 Oxygen, 7.96 Hydrogen and 6.51 Nitrogen. It has a boiling point (K) at 1112.08, a melting point (K) 718.75, a critical temperature (K) 1018.37 and a critical pressure of 13.42. Gopten has a critical volume of 1265.5. It has a Gibbs energy value of -234.71 and a Log P value of 2.9. Gopten has a 117.2 MR cm^3/mol. It has a Henry's law value of 17.62 and it has a Heat of form value of -858.91, Gopten also has a tSPA value of 95.94, a C Log P value of 1.05352 and a CMR value of 11.8329. This API has an

ACD/Log P of 4.9, an ACD/Log D (pH5.5) value 2.64, an ACD/BCF (pH5.5) value of 17, and an ACD/KOC (pH5.5) value of 60.14 and 7 H bond acceptors. There are 10 freely rotating bonds and it has an Index of Refraction of 1.549 and a Molar Volume of 364.551. The surface tension associated with it is 48.73; Gopten has a flash point of 332.42 °c, and a boiling point of 626.044°c. Gopten has an ACD/Log D (pH7.4) of 1.33, an ACD/BCF (pH7.4) of 1, an ACD/KOC (pH7.4) of 2.99 and it has the ability to donate 2 H bonds. The polar surface area of Gopten is 95.94. The molar refractivity of Gopten is 116.03; it also has a polarizability value of 45.998, a density of 1.181 and a value for enthalpy of vaporisation of 97.42 and a vapour pressure of 5.57E-14.

Quinapril

Quinapril is an API which has physicochemical properties including an exact mass of 438.22 and a molecular weight of 438.52. There are no Fluorine, Sulphur or Chlorine elements present but there are 68.47 Carbon, 18.24 Oxygen, 6.9 Hydrogen and 6.39 Nitrogen. It has a boiling point (K) at 1156.01, a melting point (K) 762.3, a critical temperature (K) 1049.32 and a critical pressure of 14.17. Quinapril has a critical volume of 1273.5. It has a Gibbs energy value of -152.35 and a Log P value of 3.17. Quinapril has a 121.81cm^3/mol MR value. It has a Henry's law value of 19.06 and it has a Heat of form value of -694.63, Quinapril also has a tSPA value of 95.94, a C Log P value of 1.9111 and a CMR value of 12.2025. This API has an ACD/Log P of 4.788, an ACD/Log D (pH5.5) value 2.6, an ACD/BCF (pH5.5) value of 16.7, and an ACD/KOC (pH5.5) value of 62.44 and 7 H bond acceptors. There are 10 freely rotating bonds and it has an Index of Refraction of 1.578 and a Molar Volume of 360.125. The surface tension associated with it is 52.295; Quinapril has a flash point of 35.149 °c, and a boiling point of 661.974°c. Quinapril has an ACD/Log D (pH7.4) of 1.24, an ACD/BCF (pH7.4) of 1, an ACD/KOC (pH7.4) of 2.68 and it has the ability to donate 2 H bonds. The polar surface area of Quinapril is 95.94. The molar refractivity of Quinapril is 119.511; it also has a polarizability value of 47.378, a density of 1.218 and a value for enthalpy of vaporisation of 102.322 and a vapour pressure of 3.24E+02.

Halobetasol

Halobetasol is an API which has physicochemical properties including an exact mass of 470.93 and a molecular weight of 470.93. There are no Nitrogen or Sulphur elements present but there are 61.21 Carbon, 16.99 Oxygen, 8.07 Fluorine, 6.21 Hydrogen and Chlorine 7.53. It has a boiling point (K) at 1081.73, a melting point (K) 758.26, a critical temperature (K) 957.88 and a critical pressure of 13.11. Halobetasol has a critical volume of 1277.5. It has a

Gibbs energy value of -764.67 and a Log P value of 2.08. Halobetasol has a 116.59cm^3/mol MR value. It has a Henry's law value of 12.7 and it has a Heat of form value of -1336.53, Halobetasol also has a tSPA value of 80.67, a C Log P value of 1.9538 and a CMR value of 11.6359. This API has an ACD/Log P of 2.947, an ACD/Log D (pH5.5) value 2.95, an ACD/BCF (pH5.5) value of 102.28, and an ACD/KOC (pH5.5) value of 955.48 and 5 H bond acceptors. There are 6 freely rotating bonds and it has an Index of Refraction of 1.551 and a Molar Volume of 369.639. The surface tension associated with it is 47.439; Halobetasol has a flash point of 298.944 °c, and a boiling point of 570.691°c. Halobetasol has an ACD/Log D (pH7.4) of 2.95, an ACD/BCF (pH7.4) of 102.28, an ACD/KOC (pH7.4) of 955.48 and it has the ability to donate 1 H bond. The polar surface area of Halobetasol is 80.67. The molar refractivity of Halobetasol is 117.848; it also has a polarizability value of 46.718, a density of 1.312 and a value for enthalpy of vaporisation of 98.309 and a vapour pressure of 0.

Momentasome fuorate monohydrate

Momentasome fuorate monohydrate is an API which has physicochemical properties including an exact mass of 520.14 and a molecular weight of 521.43. There are no Fluorine, Nitrogen or Sulphur elements present but there are 62.19 Carbon, 18.4 Oxygen, 5.8 Hydrogen and Chlorine 13.6. It has a boiling point (K) at 1248.4, a melting point (K) 877.28, a critical temperature (K) 1957.14 and a critical pressure of 13.94. Momentasome fuorate monohydrate has a critical volume of 1386.5. It has a Gibbs energy value of -357.68 and a Log P value of 3.21. Momentasome fuorate monohydrate has a 133.63cm^3/mol MR value. It has a Henry's law value of 14.45 and it has a Heat of form value of -954.88, Momentasome fuorate monohydrate also has a tSPA value of 89.9, a C Log P value of 2.37052 and a CMR value of 13.3576. This API has an ACD/Log P of 2.675, an ACD/Log D (pH5.5) value 2.68, an ACD/BCF (pH5.5) value of 63.48, and an ACD/KOC (pH5.5) value of 679.14 and 4 H bond acceptors. There are 4 freely rotating bonds and it has an Index of Refraction of 1.6 and a Molar Volume of 316.548. The surface tension associated with it is 55.51599; Momentasome fuorate monohydrate has a flash point of 308.544 °c, and a boiling point of 586.566°c. Momentasome fuorate monohydrate has an ACD/Log D (pH7.4) of 2.68, an ACD/BCF (pH7.4) of 63.48, an ACD/KOC (pH7.4) of 679.12 and it has the ability to donate 2 H bonds. The polar surface area of Momentasome fuorate monohydrate is 74.6. The molar refractivity of Momentasome fuorate monohydrate is 108.251; it also has a polarizability value of 42.914, a density of 1.35 and a value for enthalpy of vaporisation of 100.559 and a vapour pressure of 0.

Clobetasol propionate

Clobetasol propionate is an API which has physicochemical properties including an exact mass of 466.19 and a molecular weight of 466.97. There are no Nitrogen or Sulphur elements present but there are 64.3 Carbon, 17.13 Oxygen, 6.91 Hydrogen, Fluorine 4.07 and Chlorine 7.59. It has a boiling point (K) at 1114.28, a melting point (K), 769.66, a critical temperature (K) 976.38 and a critical pressure of 13.18. Clobetasol propionate has a critical volume of 1308.5. It has a Gibbs energy value of -565.83 and a Log P value of 2.63. Clobetasol propionate has a 121.17cm^3/mol MR value. It has a Henry's law value of 12.87 and it has a Heat of form value of -11146.9, Clobetasol propionate also has a tSPA value of 80.67, a C Log P value of 3.15848 and a CMR value of 12.0842. This API has an ACD/Log P of 3.142, an ACD/Log D (pH5.5) value 3.14, an ACD/BCF (pH5.5) value of 143.87, and an ACD/KOC (pH5.5) value of 1219.77 and 5 H bond acceptors. There are 6 freely rotating bonds and it has an Index of Refraction of 1.56 and a Molar Volume of 364.135. The surface tension associated with it is 48.914; Clobetasol propionate has a flash point of 297.905 °c, and a boiling point of 568.973°c. Clobetasol propionate has an ACD/Log D (pH7.4) of 3.14, an ACD/BCF (pH7.4) of 143.87, an ACD/KOC (pH7.4) of 1219.77 and it has the ability to donate 1 H bond. The polar surface area of Clobetasol propionate is 80.67. The molar refractivity of Clobetasol propionate is 117.751; it also has a polarizability value of 46.68, a density of 1.282 and a value for enthalpy of vaporisation of 98.067 and a vapour pressure of 0.

Dexamethasome dipropionate

Dexamethasome dipropionate is an API which has physicochemical properties including an exact mass of 504.25 and a molecular weight of 504.59. There are no Chlorine, Nitrogen or Sulphur elements present but there are 66.65 Carbon, 22.2 Oxygen, 7.39 Hydrogen and 3.77 Fluorine. It has a boiling point (K) at 1203.71, a melting point (K), 815.88, a critical temperature (K) 1013.35 and a critical pressure of 11.1. Dexamethasome dipropionate has a critical volume of 1453.5. It has a Gibbs energy value of -839.01 and a Log P value of 2.49. Dexamethasome dipropionate has a 132.14cm^3/mol MR value. It has a Henry's law value of 14.92 and it has a Heat of form value of -1510.34, Dexamethasome dipropionate also has a tSPA value of 106.97, a C Log P value of 2.26712 and a CMR value of 13.173. This API has an ACD/Log P of 3.666, an ACD/Log D (pH5.5) value 3.67, an ACD/BCF (pH5.5) value of 360.11, and an ACD/KOC (pH5.5) value of 2352.34 and 7 H bond acceptors. There are 9 freely rotating bonds and it has an Index of Refraction of 1.55 and a Molar Volume of 403.95.

The surface tension associated with it is 49.18999; Dexamethasome dipropionate has a flash point of 318.585 °c, and a boiling point of 603.169°c. Dexamethasome dipropionate has an ACD/Log D (pH7.4) of 3.67, an ACD/BCF (pH7.4) of 360.11, an ACD/KOC (pH7.4) of 2352.33 and it has the ability to donate 1 H bond. The polar surface area of Dexamethasome dipropionate is 106.97. The molar refractivity of Dexamethasome dipropionate is 128.667; it also has a polarizability value of 51.008, a density of 1.249 and a value for enthalpy of vaporisation of 102.93 and a vapour pressure of 0.

Hytrin

Hytrin is an API which has physicochemical properties including an exact mass of 387.19 and a molecular weight of 387.43. There are no Fluorine, Chlorine or Sulphur elements present but there are 58.9Carbon, 16.52 Oxygen, 6.5 Hydrogen and 18.08 Nitrogen. It has a boiling point (K) at 1075.94, a melting point (K), 854.39, a critical temperature (K) 1034.08 and a critical pressure of 23.05. Hytrin has a critical volume of 1030.5. It has a Gibbs energy value of 399.62 and a Log P value of 1.13. Hytrin has a 107.91cm^3/mol MR value. It has a Henry's law value of 22.02 and it has a Heat of form value of -180.9, Hytrin also has a tSPA value of 101.98, a C Log P value of 2.18152 and a CMR value of 10.3025. This API has an ACD/Log P of 0.797, an ACD/Log D (pH5.5) value -0.25, an ACD/BCF (pH5.5) value of 1, and an ACD/KOC (pH5.5) value of 5.77 and 9 H bond acceptors. There are 4 freely rotating bonds and it has an Index of Refraction of 1.636 and a Molar Volume of 290.672. The surface tension associated with it is 64.138; Hytrin has a flash point of 355.66 °c, and a boiling point of 664.477°c. Hytrin has an ACD/Log D (pH7.4) of 0.74, an ACD/BCF (pH7.4) of 2.1, an ACD/KOC (pH7.4) of 57.08 and it has the ability to donate 2 H bond. The polar surface area of Hytrin is 103.04. The molar refractivity of Hytrin is 104.264; it also has a polarizability value of 1.334, a density of 1.333 and a value for enthalpy of vaporisation of 97.721 and a vapour pressure of 1.67E-12.

Androgel

Androgel is an API which has physicochemical properties including an exact mass of 288.21 and a molecular weight of 288.42. There are no Fluorine, Chlorine, Nitrogen or Sulphur elements present but there are 79.12 Carbon, 11.09 Oxygen and 9.78 Hydrogen. It has a boiling point (K) at 837.91, a melting point (K), 539.19, a critical temperature (K) 850.34 and a critical pressure of 20.2. Androgel has a critical volume of 89.5. It has a Gibbs energy value of 26.12 and a Log P value of 3.31. Androgel has an 84.29cm^3/mol MR value. It has a Henry's law value of 6.84 and it has a Heat of form value of -428.91, Androgel also has a

tSPA value of 37.3, a C Log P value of -0.11016 and a CMR value of 8.5194. This API has an ACD/Log P of 3.179, an ACD/Log D (pH5.5) value 1.38, an ACD/BCF (pH5.5) value of 153.38, and an ACD/KOC (pH5.5) value of 1276.96 and 2 H bond acceptors. There is 1 freely rotating bond and it has an Index of Refraction of 1.56 and a Molar Volume of 256.96. The surface tension associated with it is 44.49; Androgel has a flash point of 184.655 °c, and a boiling point of 432.925°c. Androgel has an ACD/Log D (pH7.4) of 3.18, an ACD/BCF (pH7.4) of 153.38, an ACD/KOC (pH7.4) of 1276.96 and it has the ability to donate 1 H bond. The polar surface area of Androgel is 37.3. The molar refractivity of Androgel is 83.113; it also has a polarizability value of 32.949, a density of 1.122 and a value for enthalpy of vaporisation of 79.521 and a vapour pressure of 1.71E-08.

Marcaine

Marcaine is an API which has physicochemical properties including an exact mass of 288.22 and a molecular weight of 288.43. There are no Fluorine, Chlorine or Sulphur elements present but there are 74.96 Carbon, 5.55 Oxygen, 9.78 Hydrogen and 9.71 Nitrogen. It has a boiling point (K) at 800.71, a melting point (K), 553.16, a critical temperature (K) 934.64 and a critical pressure of 18. Marcaine has a critical volume of 922.5. It has a Gibbs energy value of 282.08 and a Log P value of 3.86. Marcaine has an MR value of 89.94cm^3/mol. It has a Henry's law value of 9.43 and it has a Heat of form value of -185.46, Marcaine also has a tSPA value of 32.34, a C Log P value of 3.6912 and a CMR value of 8.8499. This API has an ACD/Log P of 3.312, an ACD/Log D (pH5.5) value 1.27, an ACD/BCF (pH5.5) value of 1.77, and an ACD/KOC (pH5.5) value of 13.82 and 3 H bond acceptors. There are 5 freely rotating bond and it has an Index of Refraction of 1.547 and a Molar Volume of 279.243. The surface tension associated with it is 41.579; Marcaine has a flash point of 209.878 °c, and a boiling point of 423.422°c. Marcaine has an ACD/Log D (pH7.4) of 2.92, an ACD/BCF (pH7.4) of 77.82, an ACD/KOC (pH7.4) of 606.39 and it has the ability to donate 1 H bond. The polar surface area of Marcaine is 32.34. The molar refractivity of Marcaine is 88.62; it also has a polarizability value of 35.132, a density of 1.033 and a value for enthalpy of vaporisation of 67.775 and a vapour pressure of 0.

Warfarin

Warfarin is an API which has physicochemical properties including an exact mass of 308.1 and a molecular weight of 308.33. There are no Fluorine, Nitrogen, Chlorine or Sulphur elements present but there are 74.01 Carbon, 20.76 Oxygen, and 5.23 Hydrogen. It has a boiling point (K) at 912.75, a melting point (K), 558.33, a critical temperature (K), 958.8 and

a critical pressure of 23.05. Warfarin has a critical volume of 880.5. It has a Gibbs energy value of -163 and a Log P value of 2.97. Warfarin has an MR value of 85.93cm^3/mol. It has a Henry's law value of 11.03 and it has a Heat of form value of -427.13, Warfarin also has a tSPA value of 63.6, a C Log P value of 2.9013 and a CMR value of 8.7182. This API has an ACD/Log P of 3.129, an ACD/Log D (pH5.5) value 2.09, an ACD/BCF (pH5.5) value of 12.83, and an ACD/KOC (pH5.5) value of 109.47 and 4 H bond acceptors. There are 5 freely rotating bond and it has an Index of Refraction of 1.635 and a Molar Volume of 235.758. The surface tension associated with it is 58.658; Warfarin has a flash point of 188.828 °c, and a boiling point of 515.155°c. Warfarin has an ACD/Log D (pH7.4) of 0.33, an ACD/BCF (pH7.4) of 1, an ACD/KOC (pH7.4) of 1.89 and it has the ability to donate 1 H bond. The polar surface area of Warfarin is 63.6. The molar refractivity of Warfarin is 84.447; it also has a polarizability value of 33.477, a density of 1.308 and a value for enthalpy of vaporisation of 82.854 and a vapour pressure of 1.16E-07.

Physicochemical properties and values compared for the groups of API's in table IX

Group 1

In group 1 the chemicals Atenonol, Meprobamate and Gabapentin were identified. The similar physicochemical properties in the group are the fact that the chemicals all have the same elements present which are Carbon, Oxygen, Hydrogen and Nitrogen. All three chemicals have a low C Log P value ranging between -0.66 and 0.915, which is the lowest of all identified groups. This group also has low ACD/LogD (pH5.5), ACD/LogP (pH5.5), ACD/BCF (pH5.5), ACD/Log D (pH7.4) values. Additionally surface tension values are similar and H bond donor ability has a tendency to be higher in this group than the other identified groups.

Group 2

Group 2 chemicals were identified as Meperidine and Brofen. Both of these API contain no Fluorine, Sulphur or Chlorine. Meperidine has Nitrogen present but Brofen does not. The chemicals both have a similar Henry's Law value, a similar C Log P value and a similar CMR value. They have the same number of freely rotating bonds (4). They have a similar value for Index of Refraction and Surface Tension and a similar Boiling point.

Group 3

This group of API consist of 2 chemicals which are Isoflurane and Severane. These chemicals both contain no Sulphur and no Nitrogen but Isoflurane contains Chlorine. The API in this group contain the lowest Gibbs Energy and Henry's Law values identified among the data set which has been defined by PCA groupings on the score plot. Similar characteristics in this group between the two chemicals are also a low Heat of Form value and the same value for tPSA. They also have similar C Log P values, CMR values, Vapour pressure values, Enthalpy of vaporisation, Density, Polarisation value, no H bond acceptors, and similar ACD/Log D (pH7.4 values. Both API have a similar Boiling point, and Surface Tension and Molar Volume. The Index of Refraction is also similar and the number of Freely rotating bonds is the same (2). The number of H bond acceptors is the same and the ACD/ Log D value (pH5.5) and ACD/Log P values are similar.

Groups 4, 5, 7, 8 and 10

Groups 4, 5, 7, 8 and 10 only contain one chemical each. It is therefore not possible to compare the common physicochemical features in these groups.

Group 6

Group 6 contains two API's these are Calcijex and Paricalcitrol. These chemicals have very similar physicochemical characteristics. These include the same exact mass, the same molecular weight, the same number of Carbons, Oxygen and the same number of Hydrogen's. Both chemicals have no Fluorine, Sulphur, Nitrogen or Chlorine. The chemicals have similar boiling points, melting points, critical temperature values, and critical pressure values. The API have similar Log P numbers, MR values, Henry's Law values, similar Heat of Form values and the same tPSA values. Calcijex and Paricalcitrol have similar CMR values, ACD/Log P (pH5.5) and ACD/Log D values. The ACD/BCF (pH5.5) values are very similar and also higher than those of the other groups, with the exception of group 7. The ACD/KOC (pH5.5) values are a lot higher in this group than in all other groups. Both chemicals have 3 H bond acceptors and a higher number of freely rotating bonds than most other groups identified. The API's have similar molar volumes and boiling points and the same flash point values. High ACD/BCF (pH7.4) and ACD/KOC (pH7.4) values are indicative of this group. The group of chemicals also has the same number of H bond donors, the same polar surface volume, similar molar refractivity values, similar polarizability values and Enthalpy of vaporisation values.

Group 9

Group 9 is the largest group of chemicals identified. It contains chemicals Gopten, Quinapril, Halobetasol, Mometasone furoate monohydrate, Clobetasol propionate and Dexamethasone dipropionate. The chemicals in this group are identified by the fact that they have a very similar number of Carbons and a similar number of Oxygen and Hydrogen atoms. None of the chemicals in this group have Sulphur present but the amount of Fluorine, Nitrogen and Chlorine varies between the chemicals. The API's have a similar boiling point which is higher in this group than in the other groups except for group 10. The chemicals in this group also have similar critical pressure values, critical pressure values, and critical volume values, Gibbs Energy values and Log P values. Heat of form values in this group are similar and all of these values are negative. The chemicals have similar tPSA values and CMR values

Group 11

In group 11 there are three chemicals. These are Warfarin, Marcaine and Androgel. Similar physicochemical characteristics between the three chemicals are the amount Carbon present. All of these chemicals have no Fluorine, Sulphur, and Chlorine. One chemical, Marcaine has Nitrogen present. The chemicals have similar melting points, MR values, CMR values, ACD/Log P values and molar volumes. The API's are all able to donate one H bond and they have a similar molar refractivity value.

Analysis of grouping in table IX based on physicochemical property values against average values for the data set.

Information relating to each identified group is given below and this indicated whether the values given for each physicochemical characteristic was below or above the average value for the data set. Using this information it was possible to define the following identifying features in each group.

Group 1

Group one features API's that had the following characteristics which were below the average value for the data set. Elements C, F, H, S and Cl were lower than the average for the data set as were molecular weight and the exact mass. Both the boiling point and melting point were lower than the average value. Critical temperature, critical volume, Log P, MR, CLogP and CMR were lower than the averages for the physicochemical characteristics within the data set.

Characteristics ACD/Log P (pH5.5), ACD/LogD (pH5.5), ACD/BCF (pH5.5) and ACD/KOC (pH5.5) were found to be lower than the average values for the data set. The values for H bond acceptors; molar volume, surface tension, flash point [K] and boiling point [K] were lower than the average values for the data set. Characteristics ACD/BCF (pH7.4), ACD/KOC (pH7.4), Molar Refractivity, Polarizability, Density and Enthalpy of vaporisation were lower than average values in group one.

Higher than the average data set values in group one included atoms of the elements O and N. Values for Critical pressure, Heat of form and H bond donor were found to be higher in this group than the average values.

Variability is found in data for Gibbs energy, Henry's Law, tPSA, freely rotating bonds, Index of refraction, ACD/LogD (7.4), Polar surface area and vapour pressure within group one. These physicochemical characteristics were both below and above the average values for this particular set of data. This could indicate that these physicochemical characteristics were not of importance in this group.

Group 2

Group two was characterised by the following physicochemical features that were lower than average for the data set; exact mass, Molecular weight and elements F, S, N and Cl. The boiling point and melting point, ACD/KOC (pH7.4), H bond donors, critical temperature and critical volume were below the averages for the dataset. Log P values, polar surface area, molar refractivity and polarizability were also below average values within this data set. The values for MR, Henry's Law, tPSA, CMR and Enthalpy of vaporisation were below average values. Characteristics ACD/BCF (pH5.5), ACD/KOC (pH5.5), ACD/BCF (pH7.4), availability of H bond acceptors and number of freely rotating bonds were also below the average value for the data set.

Physicochemical characteristics with values higher than average were elements C and H and features critical pressure, Gibbs energy, Heat of Form, C LogP and ACD/LogD (pH7.4).

Variability in group two was found in the data associated with physicochemical characteristics ACD/Log P (pH5.5), density, vapour pressure and the element O.

Group 3

Lower than average values for the data set were found in group 3 for the elements C, O, H, S and N. Both the exact mass of the API and the molecular weight were lower than average in

group three. The following physicochemical characteristics were also found to be lower than average within the data set; critical volume and Gibbs energy, the boiling point and the melting point, critical temperature, the Log P value, MR, Henry's Law and heat of form. The values for tPSA, C LogP, CMR, ACD/BCF (pH5.5), ACD/KOC (pH5.5), number of H bond acceptors, number of freely rotating bonds and Index of Refraction were lower than the average of the data set. Features including Molar volume, surface tension, flash point and boiling point[k], ACD/KOC (pH7.4), the number of H bond donors, polar surface area and molar refractivity, enthalpy of vaporisation, density and polarizability were all below the average values for the data set.

Group three had the following common physicochemical features which were above average in the data set; the element F, critical pressure, the values for both ACD/Log D (pH5.5) and ACD/Log D (pH7.4) and vapour pressure.

Physicochemical features which vary and occur both above and below the value for the average in the data set include the element Chlorine and the value for ACD/Log P (pH5.5).

Group 4

Group four contains only one API. This group was differentiated from the other groups identified by the following features identified as below the average values in the data set; exact mass, molecular weight, the elements O, F, S, N, and Cl, both the boiling point and the melting point, critical pressure, critical temperature and critical volume, Log P, MR and Henry's Law. Other physicochemical characteristics below the average of the data set include tPSA, C Log P, CMR, ACD/BCF (pH5.5), ACD/KOC (pH5.5), and number of H bond acceptors along with the number of freely rotating bonds.

Higher than average physicochemical values were found for elements C and H and the characteristics Gibbs energy, Heat of Form, ACD/LogP (pH5.5), ACD/LogD (pH5.5), Index of refraction and ACD/LogD (pH7.4).

Group 5

There was only one API in this group, it was differentiated from the other groups by the following features; physicochemical characteristics below average values in the data set include exact mass, molecular weight and elements C, O, F, H, S and N, boiling point, critical temperature and critical pressure. Gibbs energy, Log P, Henry's Law, tPSA, number of H bond acceptors and number of freely rotating bonds were also below average values in this

group. Values for molar volume, surface tension, flash point [K] and boiling point [K], number of H bond donors, polar surface area, molar refractivity, polarizability, density, enthalpy of vaporization and vapour pressure were also below average values in group five.

Physicochemical characteristics above average values in the data set were as follows; the element Cl, melting point, critical volume, MR, Heat of Form, C Log P, CMR, ACD/Log P (pH 5.5), ACD/Log D (pH5.5) and ACD/BCF (pH5.5). Other features with values higher than the average value were ACD/KOC (pH5.5), Index of refraction, ACD/Log D (pH7.4), ACD/BCF (pH7.4) and ACD/KOC (pH7.4).

Group 6

Group six had the following physicochemical features in common which were lower than the average value for the data set; exact mass, molecular weight, the elements O, F, S, N, and Cl, critical pressure, Log P, Henry's Law, tPSA, number of H bond acceptors. The values for flash point, boiling point, polar surface area, density and vapour pressure were also lower than the average value for the data set.

Physicochemical characteristics which had values higher than the average in the data set include the elements C and H, boiling point, melting point, critical temperature, critical volume, Gibbs energy, MR, heat of form and C Log P. Other features which were of a higher than average value were CMR, ACD/Log P (pH5.5), ACD/Log D (pH5.5), ACD/BCF (pH5.5), ACD/KOC (pH5.5), number of freely rotating bonds, index of refraction and molar volume. Characteristics including ACD/BCF (pH7.4), ACD/Log D (7.4), ACD/KOC (7.4), number of H bond donors, molar refraction, polarizability and enthalpy of vaporisation also had greater than the average values.

Within group six there was only one characteristic which was variable; it does not appear to be either lower or higher than the average, this was surface tension.

Group 7

There was only one API in group seven the characteristics of this included physicochemical features which were below the average value for the data set. These were elements F, H, S, N and Cl, critical pressure, Gibbs energy, Log P, Heat of form, flash point, the number of H bond donors, density and vapour pressure. All other physicochemical characteristics were above the average values for the data set.

Group 8

There was only one API in group eight and the characteristics of this included the following physicochemical features which were below average in the data set; the elements C, O, H, N and Cl, critical pressure, Gibbs energy, Log P, Henry's Law, Heat of Form, ACD/BCF (pH5.5), ACD/KOC (pH5.5), the number of H bond acceptors and surface tension. This group also included flash point [k] and boiling point [k], ACD/BCF (pH7.4) and the number of H bond donors. The values for polar density and vapour pressure were unknown. All other physicochemical properties in this group were higher than the given average value.

Group 9

Group nine was the largest of the groups identified and it had the following below average physicochemical characteristics within the data set; the elements S and N, critical pressure, Log P, ACD/BCF (pH5.5), ACD/KOC (pH5.5), ACD/BCF (pH7.4), ACD/KOC (pH7.4), number of H bond donors and density.

Common physicochemical characteristics which were above the average for the data set were; boiling point, melting point, critical temperature, critical volume, MR, Henry's Law, tPSA and CMR. Other characteristics above average include ACD/Log P (pH5.5), ACD/Log D (pH5.5), molar refractivity, enthalpy of vaporisation and polarizability.

There were a number of physicochemical characteristics which were variable either below or above the average in the group. These included the following characteristics exact mass, molecular weight, elements C, O, F, H and Cl, Gibbs energy, heat of form, C Log P, number of H bond acceptors and number of freely rotating bonds. Also included in this category were surface tension, flash point [K], boiling point [K], polar surface area and vapour pressure.

Group 10

There was only one API in group ten. This API was identified by the following characteristics which were lower than the average value in the data set; exact mass, molecular weight and the elements O, F, H, S and Cl. This group also featured Log P, ACD/Log D (pH5.5), ACD/BCF (pH5.5), ACD/KOC (pH5.5), it had a lower than average number of freely rotating bonds and molar volume. Other lower than average value physicochemical characteristics were ACD/BCF (pH7.4), ACD/KOC (pH7.4), number of H bond donors, molar refractivity, polarizability, density and enthalpy of vaporisation. All other physicochemical features had values greater than the averages determined for the data set.

Group 11

Group eleven was composed of several API's, these had the following physicochemical features which were below the average value within the data set; exact mass, molecular weight, the elements F, S and Cl, both boiling point and melting point, critical volume and Log P. Other characteristics included MR, tPSA, CMR, ACD/BCF (pH5.5), ACD/KOC (pH5.5), number of H bond acceptors, number of freely rotational bonds, molar volume, flash point [K] and boiling point [K]. Physicochemical features below the average value for the data set also included ACD/BCF (pH7.4), ACD/KOC (pH7.4), number of H bond donors, polar surface area, molar refractivity, polarizability, density, enthalpy of vaporisation and vapour pressure.

In group eleven there were several physicochemical characteristics which had common values greater than the average. These were the element C, Gibbs energy, Heat of Form, ACD/Log P (pH5.5), ACD/Log D (pH5.5), Index of refraction and ACD/Log D (pH7.4).

There were a number of physicochemical characteristics which were variable within group eleven. These had values both above and below the average within the group. These included the elements O, H and N, critical temperature, critical pressure, Henry's Law, C Log P and surface tension.

A flow chart was constructed to help determine possible distinguishing characteristics for each group (figure XI).

Figure XI indicates the simplest way to distinguish identified groups from each other. Using this flow chart it is possible to determine which potential group an API may belong to within this data set. The flow chart is primarily based on elements identified in the API in each group for this particular data set. The flow chart also indicates key physicochemical characteristics that can be used to define each group. These can be listed as molecular weight, elements present, the boiling point, number of H bonds which can be donated, ACD/Log D (pH7.4) and ACD/BCF (pH5.5).

Figure XI Distinguishing groupings simplified

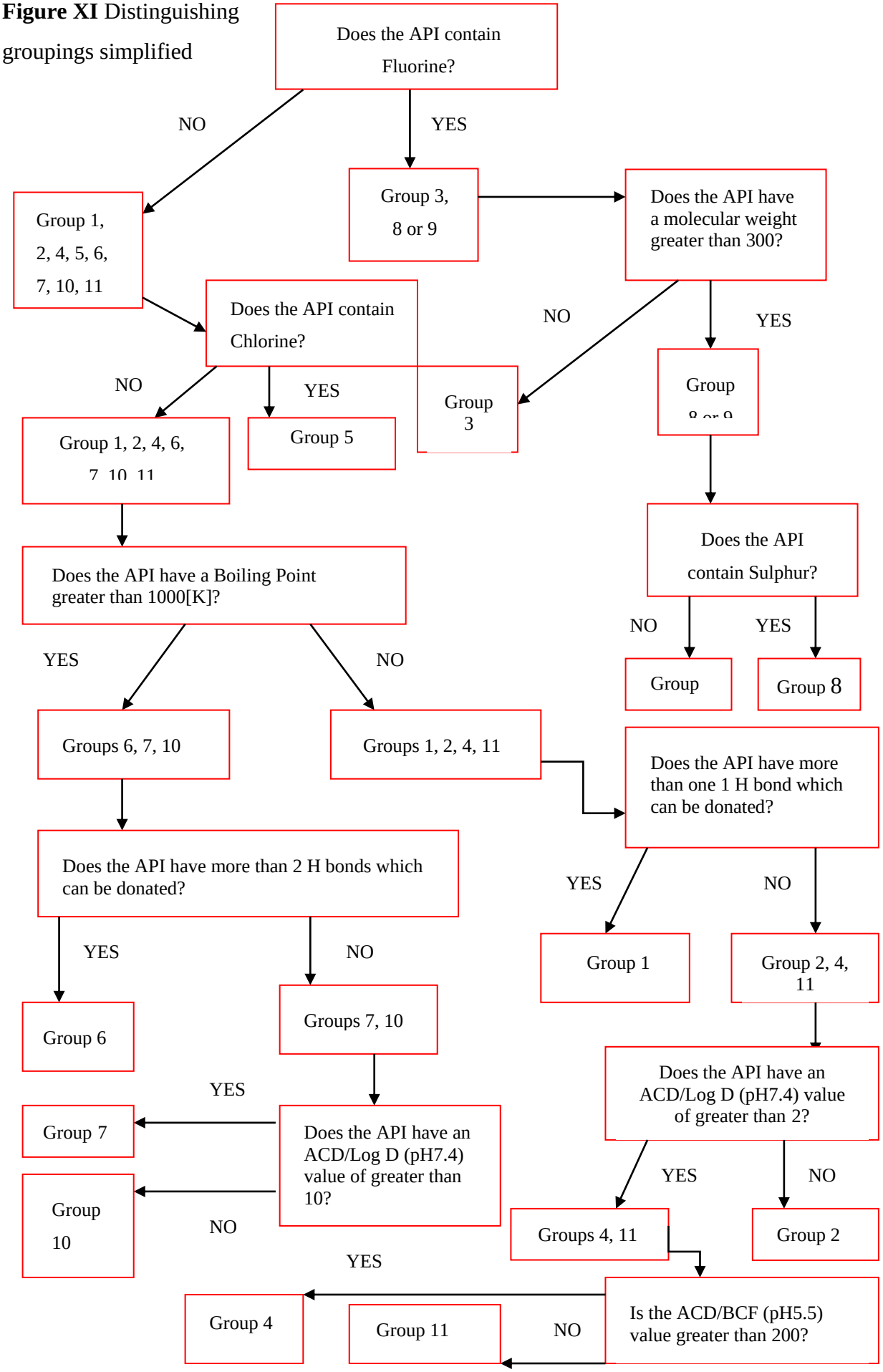


Table XII

Variable name **Functional and structural features**	**Principal component associated with the variable**	**Variable name** **Physicochemical features**	**Principal component associated with the variable**
Primary	c7 c8 c9	Dermatological classification	c10 c13
Secondary	c7 c11 c12	Nasal and inhalation classification	c4 c5
Tertiary	c11 c13	Injectable classification	c4 c5
Aromatic/enamine	c2 c9 c12	Antibiotic classification	c4 c5
Primary 1	c6 c8	API classification	c4 c5
Secondary 1	c8 c14	Exact mass	c2 c6
Tertiary 1	c5 c6	Molecular weight	c6
Vinyl alcohol	c10	Contains C	c8 c12
Phenol	c7	Contains O	c8 c12 c13
Carboxylic acid	c8 c11 c13	Contains F	c13
Ketone	c3 c10 c12	Contains S	c9 c11
Ester	c12	Contains N	c2 c3
1 amide	c10	Contains P	c4 c11 c13
2 amide	c2 c6 c13	Contains Na	c4
Tertiary amide	c2 c12 c13	Contains I	c6 c8
Thioester	c11	Contains Cl	c8 c12

Table XII

Variable name **Functional and structural features**	**Principal component associated with the variable**	**Variable name** **Physicochemical features**	**Principal component associated with the variable**
Oxime	c10	Boiling Point [K]	c1
Urea	c11 c13	Melting Point [K]	c1 c3
Guanidine	c12	Critical Temperature [K]	c1 c3
Ether	c6 c8 c10	Critical Pressure [Bar]	c3
Thioether	c6 c9	Critical Volume (cm3/mol)	c1
Fluorine	c5 c7 c11	Gibbs Energy (KJ/mol)	c5
Pyridine	c4	Log P	c11 c13
Alkyl halide	c10 c12	MR (cm3/mol)	c1
Aryl halide	c4	Henrys Law	c1 c3
Alkenes	c5 c7	tPSA	c3
Phosphonate	c4 c13	C Log P	c1
Hydrozone	c4 c11	CMR	c1 c3
Other features	c4	ACD/Log P	c1
Phosphate	c4	ACD/Log D (ph5.5)	c1 c3 c13
Nitro	c6	ACD/BCF (pH5.5)	c1 c5 c7 c9
Nitrate	c14	ACD/KOC (pH5.5)	c1 c5 c7 c9
Steroid	c3 c5 c12	H bond acceptors	c2 c3

Table XII

Variable name **Functional and structural features**	**Principal component associated with the variable**	**Variable name** **Physicochemical features**	**Principal component associated with the variable**
O-heterocyclic	c14	Freely rotating bonds	c2
N-heterocyclic	c12 c14	Index of Refraction	c1 c6
S-heterocyclic	c6 c9	Molar Volume (cm)	c2 c7
Long alkyl	c7	Surface Tension dyne/cm	c6
Phenyl ring	c8 c9	Flash Point	c2
Erythromycin derivative	c10	Boiling Point (°c)	c2
Tetracycline	c10 c12	ACD/BCF (pH7.4)	c1 c5 c7 c9
Macrocyclic	c13	ACD/KOC (pH7.4)	c1 c5 c7 c9
Macrolide	c7 c10	H bond donors	c2
Barbiturate	c11 c14	Polar surface area A	c2
Water	c10	Molar Refractivity (cm)	c2 c7
Ethanol	c10	Enthalpy of vaporisation kJ/mo	c2
HCL	c10 c12		
Gd3+	c8 c11 c13		

Table VII Eigenvalues from the first 14 principal components identified by PCA of database three. Where c stands for principal component and the number following the c is the principal component of interest.

Table XIII

Variables Functional groups and structural features	**Variables** Physicochemical properties
Enone groups	Heat of Form
Oxazolidinone groups	ACD/Log D (pH7.4)
Sulfonamide groups	Polazirability
Sulfone groups	Density
N-Oxide groups	Vapour pressure
Alkyl groups greater than 5 Carbons	
Carbamate groups	
Hormone structural features	
Na+ associations	
Hydrogen associations	
Benzodiazepine structures	

Table XIII indicates the variables which are not found within the first 14 principal components of the scree plot.

Table XIV

Product Name	Cleaning Agent	Solubility
Beclomethasone dipropionate	Acetone or DMF	Very soluble in DMF and soluble in acetone
Beclomethasone dipropionate monohydrate	Acetone or DMF	Very soluble in DMF and soluble in acetone
Fluticasone propionate	Acetone	Freely soluble in acetone
Mometasone furoate anhydrous	Acetone	Freely soluble in DMF and soluble in acetone
Mometasone furoate monohydrate	Acetone	Freely soluble in DMF and soluble in acetone
Sumatriptan Base	DMF	Not Applicable
Clobetasol propionate	Acetone or DMF	Freely soluble in acetone and DMF
Dexamethasone dipropionate	Acetone or DMF	Very soluble in DMF and freely soluble in acetone
Halobetasol	Acetone or DMF	Freely soluble in acetone
Betamethasone acetate	Acetone or DMF	Very soluble in DMF and freely soluble in acetone
Betamethasone disodium phosphate	Water	Freely soluble
Doxycycline hyclate	Methanol	Freely soluble
Doxycycline monohydrate	Methanol 1% HCL	Soluble
Roxithromycin	Methanol	Soluble
Tamsulosin	DMF	Not Applicable
Iohexol	Water	Very soluble

Table XIV Pharmaceutical products and known cleaning agents used to remove them from production equipment post manufacturing. Information provided by company D.

Appendix VI

Principal Component Analysis: the model data set and the chemicals provided for the case study

Eigenanalysis of the Correlation Matrix

Eigenvalue	4.4057	4.1336	3.3076	3.2543	3.0987	2.7237	2.3525	2.1065
Proportion	0.085	0.079	0.064	0.063	0.060	0.052	0.045	0.041
Cumulative	0.085	0.164	0.228	0.290	0.350	0.402	0.448	0.488
Eigenvalue	1.9042	1.8484	1.7535	1.5966	1.5067	1.4557	1.2033	1.1304
Proportion	0.037	0.036	0.034	0.031	0.029	0.028	0.023	0.022
Cumulative	0.525	0.560	0.594	0.625	0.654	0.682	0.705	0.727
Eigenvalue	1.1118	1.0703	0.9971	0.9673	0.9219	0.9050	0.8593	0.8324
Proportion	0.021	0.021	0.019	0.019	0.018	0.017	0.017	0.016
Cumulative	0.748	0.769	0.788	0.806	0.824	0.841	0.858	0.874
Eigenvalue	0.7553	0.7276	0.6369	0.5999	0.4982	0.4660	0.4321	0.3880
Proportion	0.015	0.014	0.012	0.012	0.010	0.009	0.008	0.007
Cumulative	0.888	0.902	0.915	0.926	0.936	0.945	0.953	0.961
Eigenvalue	0.3344	0.2937	0.2399	0.2308	0.1871	0.1634	0.1369	0.1075
Proportion	0.006	0.006	0.005	0.004	0.004	0.003	0.003	0.002
Cumulative	0.967	0.973	0.977	0.982	0.985	0.988	0.991	0.993
Eigenvalue	0.0878	0.0769	0.0657	0.0541	0.0319	0.0244	0.0151	0.0000
Proportion	0.002	0.001	0.001	0.001	0.001	0.000	0.000	0.000
Cumulative	0.995	0.996	0.998	0.999	0.999	1.000	1.000	1.000

Eigenvalue	-0.0000	-0.0000	-0.0000	-0.0000
Proportion	-0.000	-0.000	-0.000	-0.000
Cumulative	1.000	1.000	1.000	1.000

Variable	PC1	PC2	PC3	PC4	PC5	PC6	PC7
Primary	0.052	-0.043	-0.031	0.030	0.071	-0.143	0.042
Secondary	0.020	-0.152	0.248	-0.056	0.111	-0.162	-0.036
Tertiary	-0.149	-0.134	0.354	-0.110	-0.177	0.086	0.031
Aromatic/enamine	0.114	-0.091	0.004	0.252	-0.303	0.056	0.091
Primary_1	-0.009	-0.085	0.210	-0.064	0.054	-0.011	0.020
Secondary_1	-0.241	-0.064	0.303	-0.157	0.086	0.084	-0.017
Tertiary_1	-0.377	0.088	-0.059	-0.008	-0.104	-0.097	-0.013
Vinyl alcohol	-0.092	0.005	-0.122	-0.341	-0.334	-0.144	-0.036
Phenol	0.059	-0.067	0.016	0.070	0.037	-0.188	0.056
Carboxylic	-0.024	-0.133	0.436	-0.098	0.024	-0.026	0.023
Ketone	-0.035	0.255	-0.122	-0.185	0.058	0.236	0.051
Thioester	0.022	0.070	0.068	-0.010	-0.028	0.225	0.105
Oxime	-0.393	-0.046	-0.060	0.192	0.018	0.001	0.037
Oxazolidinone	0.017	0.002	-0.036	-0.010	0.032	0.001	0.009
Urea	0.040	0.013	-0.022	0.064	-0.019	-0.058	0.065
Guanidine	0.062	-0.065	-0.080	0.122	-0.122	-0.120	0.137
Ether	-0.264	-0.055	-0.111	0.130	0.045	-0.003	-0.039
Sulfonamide	0.044	-0.057	0.019	0.030	0.046	-0.116	-0.060
Sulfone	-0.018	-0.019	-0.049	0.045	0.015	-0.010	0.003
N-Oxide	0.008	-0.026	-0.045	0.020	0.028	0.053	-0.579
Thioether	0.067	-0.034	0.114	0.206	-0.341	0.268	-0.027
Fluorine	0.015	0.188	-0.007	-0.063	0.080	0.221	0.099
Pyridine	0.006	0.216	0.076	0.084	-0.047	-0.118	-0.013
Alkyl halide	0.007	0.111	-0.101	-0.125	0.114	0.261	0.088
Aryl halide	0.055	0.057	0.024	0.092	0.039	-0.251	0.003
Alkene	-0.013	-0.015	-0.026	0.009	0.023	-0.059	0.002
Alkylgreater than5 C	0.050	-0.025	-0.051	0.032	0.037	-0.063	0.038
Phosphonate	-0.008	0.321	0.130	0.106	-0.073	-0.143	-0.033
Hydrozone	-0.012	0.318	0.196	0.105	-0.122	-0.046	-0.034
Other_1	-0.012	0.279	0.107	0.066	-0.035	-0.113	-0.049
Phosphate	-0.035	0.409	0.165	0.091	-0.068	-0.151	-0.068
Carbamate	0.026	-0.022	-0.023	-0.005	0.042	-0.020	-0.011

Nitro	0.068	-0.068	0.061	0.211	-0.342	0.265	-0.047
Nitrate	0.008	-0.026	-0.045	0.020	0.028	0.053	-0.579
Steroid	0.001	0.277	-0.067	-0.124	0.134	0.297	0.106
Hormone	0.042	0.009	-0.030	0.025	0.070	-0.121	0.041
O-heterocyclic	0.057	-0.054	-0.032	0.059	-0.049	0.054	-0.414
N-heterocyclic	0.123	-0.070	-0.006	0.209	-0.144	-0.118	0.157
S-heterocyclic	0.054	-0.053	0.053	0.178	-0.288	0.218	0.014
Long alkyl	0.014	-0.032	-0.007	-0.015	0.066	-0.036	-0.053
Phenyl ring	0.125	-0.101	-0.044	0.119	0.021	-0.279	0.047
Erythromycin deriv	-0.393	-0.046	-0.060	0.192	0.018	0.001	0.037
Tetracycline	-0.393	-0.046	-0.060	0.192	0.018	0.001	0.037
Macrocyclic	-0.099	0.008	-0.115	-0.346	-0.339	-0.142	-0.020
Macrolide	-0.351	-0.040	-0.053	0.119	0.033	0.024	0.006
Benzodiazepine	0.009	-0.014	0.047	-0.004	-0.033	0.077	0.066
Barbiturate	0.032	-0.007	-0.033	0.042	-0.004	-0.040	0.062
Water	-0.023	-0.005	-0.124	-0.139	-0.201	-0.079	0.073
Ethanol	-0.043	-0.009	-0.056	-0.185	-0.175	-0.071	-0.019
HCl	-0.081	0.006	-0.096	-0.298	-0.292	-0.127	-0.024
Na+	-0.029	0.383	0.158	0.085	-0.059	-0.146	-0.068
Gd3+	-0.070	-0.129	0.442	-0.154	0.036	0.017	0.008

Variable	PC8	PC9	PC10	PC11	PC12	PC13	PC14
Primary	0.125	0.289	0.015	-0.067	-0.119	0.185	-0.493
Secondary	0.093	0.046	0.048	0.089	0.298	0.150	0.090
Tertiary	-0.008	-0.006	0.163	0.014	-0.030	-0.017	-0.095
Aromatic/enamine	0.050	-0.215	-0.148	-0.004	-0.140	0.108	-0.010
Primary_1	-0.084	-0.081	-0.238	-0.085	-0.199	-0.004	0.148
Secondary_1	-0.010	-0.133	-0.188	0.003	-0.093	-0.065	-0.012
Tertiary_1	0.038	0.085	-0.068	-0.056	-0.041	-0.083	0.007
Vinyl alcohol	0.022	0.023	0.031	0.062	-0.006	0.018	0.023
Phenol	0.191	-0.101	-0.090	0.146	-0.278	-0.153	0.159
Carboxylic	0.020	-0.084	0.024	0.091	-0.042	0.017	-0.103
Ketone	0.092	-0.111	-0.119	0.135	0.044	0.035	-0.062
Thioester	-0.020	0.030	0.536	-0.047	-0.234	-0.089	0.117
Oxime	0.031	0.044	0.058	0.113	-0.054	0.221	0.075
Oxazolidinone	0.016	0.135	-0.021	-0.058	-0.001	0.116	-0.291
Urea	-0.561	0.093	-0.078	0.344	-0.072	-0.032	-0.020
Guanidine	0.004	-0.486	-0.038	-0.104	-0.155	0.165	-0.002
Ether	-0.031	-0.239	0.015	-0.122	0.171	-0.408	-0.182
Sulfonamide	0.066	0.037	0.147	0.119	0.410	0.061	0.191
Sulfone	-0.074	-0.121	0.044	-0.153	0.201	-0.404	-0.197
N-Oxide	-0.065	-0.149	0.083	0.042	-0.173	0.053	-0.110
Thioether	0.099	0.142	-0.001	0.060	0.029	-0.080	0.002
Fluorine	0.059	-0.029	0.210	0.029	-0.135	0.074	0.074
Pyridine	-0.054	-0.028	0.009	-0.108	0.044	0.058	-0.048
Alkyl halide	0.070	-0.166	-0.155	0.184	0.081	0.097	-0.058
Aryl halide	0.285	0.044	0.063	0.359	-0.126	-0.227	-0.091
Alkene	0.034	0.164	-0.138	-0.152	-0.235	-0.152	0.182
Alkylgreater than5 C	0.020	0.193	-0.080	-0.268	-0.225	0.224	-0.272
Phosphonate	-0.125	-0.026	-0.086	0.021	-0.008	0.027	0.013
Hydrozone	0.013	0.022	0.134	-0.039	-0.026	-0.045	0.018
Other_1	0.009	0.015	-0.006	-0.077	0.047	0.032	0.013
Phosphate	0.020	-0.024	-0.048	-0.031	0.030	0.015	0.001
Carbamate	-0.022	0.092	0.003	-0.069	0.163	0.057	0.132
Nitro	0.122	0.134	-0.170	0.091	0.083	-0.041	-0.015
Nitrate	-0.065	-0.149	0.083	0.042	-0.173	0.053	-0.110
Steroid	0.108	-0.146	-0.033	0.203	-0.023	0.048	-0.023
Hormone	0.263	0.040	0.032	0.384	-0.205	-0.258	-0.201
O-heterocyclic	0.071	0.011	0.010	0.112	0.080	0.069	0.160
N-heterocyclic	-0.180	-0.219	0.175	-0.042	0.023	0.030	-0.072
S-heterocyclic	0.093	0.111	-0.153	0.067	0.072	-0.055	-0.059
Long alkyl	0.033	0.132	-0.227	-0.144	-0.253	-0.174	0.403
Phenyl ring	0.187	-0.166	0.136	0.157	0.081	0.094	0.109
Erythromycin deriv	0.031	0.044	0.058	0.113	-0.054	0.221	0.075
Tetracycline	0.031	0.044	0.058	0.113	-0.054	0.221	0.075
Macrocyclic	0.014	0.025	0.027	0.051	-0.019	0.007	0.005
Macrolide	-0.007	-0.055	-0.024	-0.051	0.077	-0.226	-0.100
Benzodiazepine	-0.072	0.029	0.436	-0.096	-0.172	-0.093	0.069
Barbiturate	-0.526	0.109	-0.050	0.352	-0.059	-0.046	-0.031
Water	0.008	-0.326	-0.030	-0.014	-0.041	0.125	-0.032

Ethanol	-0.017	0.080	0.053	-0.004	0.027	-0.055	-0.008
HCl	0.010	0.060	0.038	0.051	-0.010	-0.017	0.007
Na+	0.013	-0.013	-0.049	-0.041	0.038	0.016	0.003
Gd3+	-0.017	-0.118	-0.056	0.072	-0.006	0.061	-0.135

Variable	PC15	PC16	PC17	PC18	PC19	PC20	PC21
Primary	0.109	-0.069	-0.154	0.036	-0.051	0.065	-0.096
Secondary	0.149	-0.060	-0.197	0.015	-0.077	0.054	-0.060
Tertiary	0.056	0.028	0.046	0.012	0.076	0.129	0.197
Aromatic/enamine	0.027	0.075	-0.074	-0.056	0.103	-0.039	0.086
Primary_1	-0.314	-0.070	-0.079	-0.090	0.355	-0.142	-0.407
Secondary_1	0.040	0.135	-0.021	-0.002	0.114	-0.041	-0.146
Tertiary_1	0.248	0.132	-0.038	0.088	0.012	0.010	-0.053
Vinyl alcohol	0.017	-0.035	-0.036	0.004	-0.066	-0.007	-0.179
Phenol	-0.135	-0.281	-0.226	-0.000	-0.093	0.130	0.071
Carboxylic	0.052	-0.096	0.081	0.001	-0.249	-0.048	0.102
Ketone	0.119	-0.174	0.036	0.086	-0.001	0.075	0.060
Thioester	0.040	-0.016	-0.166	0.040	-0.076	-0.062	-0.118
Oxime	-0.091	-0.095	0.073	-0.057	-0.075	-0.070	0.000
Oxazolidinone	-0.060	-0.249	-0.138	0.526	0.179	-0.524	0.159
Urea	0.068	0.020	-0.109	0.013	-0.006	0.030	0.008
Guanidine	0.057	0.065	-0.136	0.107	-0.012	0.022	0.129
Ether	0.047	-0.021	-0.176	-0.007	-0.003	0.080	0.039
Sulfonamide	0.273	-0.096	-0.042	-0.062	0.257	-0.032	-0.186
Sulfone	-0.089	-0.217	-0.007	-0.234	-0.317	-0.222	-0.087
N-Oxide	-0.018	0.047	0.052	-0.020	-0.017	-0.044	-0.024
Thioether	0.018	-0.056	-0.052	0.063	-0.171	-0.011	-0.068
Fluorine	0.064	0.215	-0.371	-0.308	-0.050	-0.263	-0.008
Pyridine	-0.014	-0.148	0.378	-0.139	0.011	-0.003	-0.175
Alkyl halide	0.081	-0.157	0.250	0.033	-0.025	0.160	0.082
Aryl halide	0.045	0.116	0.128	0.009	0.082	-0.054	-0.060
Alkene	0.467	0.324	0.320	0.041	-0.060	-0.308	0.115
Alkylgreater than5 C	0.137	-0.015	-0.068	-0.318	-0.081	0.366	-0.061
Phosphonate	-0.045	-0.011	-0.039	0.011	0.140	-0.008	-0.041
Hydrozone	0.020	-0.098	-0.051	0.113	-0.245	-0.027	-0.043
Other_1	-0.073	0.117	-0.121	-0.067	0.069	0.069	0.214
Phosphate	0.005	-0.008	-0.039	0.035	0.017	0.042	0.028
Carbamate	-0.457	0.417	0.098	0.308	-0.370	0.142	-0.023
Nitro	0.005	0.024	-0.041	-0.011	0.049	0.043	-0.019
Nitrate	-0.018	0.047	0.052	-0.020	-0.017	-0.044	-0.024
Steroid	0.046	0.007	0.021	-0.083	-0.032	-0.024	-0.024
Hormone	-0.162	0.074	0.117	0.026	0.022	0.065	-0.092
O-heterocyclic	0.151	-0.155	-0.111	0.030	-0.009	0.107	0.071
N-heterocyclic	0.097	-0.158	0.331	-0.093	-0.028	-0.129	-0.081
S-heterocyclic	-0.027	0.060	0.039	-0.041	0.069	-0.013	-0.085
Long alkyl	0.110	-0.381	-0.035	0.088	-0.193	0.071	0.114
Phenyl ring	0.064	0.139	-0.071	-0.058	-0.008	-0.057	0.117
Erythromycin deriv	-0.091	-0.095	0.073	-0.057	-0.075	-0.070	0.000
Tetracycline	-0.091	-0.095	0.073	-0.057	-0.075	-0.070	0.000
Macrocyclic	0.008	-0.027	-0.026	0.001	-0.065	-0.014	-0.194
Macrolide	0.087	0.114	-0.155	0.104	0.178	0.198	0.025
Benzodiazepine	-0.001	-0.112	0.194	0.257	0.325	0.343	0.011
Barbiturate	0.110	0.026	-0.113	0.024	-0.096	0.060	0.051
Water	0.126	0.052	-0.047	0.278	-0.145	0.079	-0.143
Ethanol	-0.223	-0.040	0.062	-0.258	0.184	-0.057	0.592
HCl	-0.082	-0.058	-0.003	-0.153	0.023	-0.067	0.057
Na+	-0.018	-0.008	-0.026	0.044	0.008	0.041	0.043
Gd3+	0.068	0.023	0.057	-0.013	-0.125	-0.014	0.122

Variable	PC22	PC23	PC24	PC25	PC26	PC27	PC28
Primary	0.081	0.116	0.169	-0.070	-0.006	-0.267	0.021
Secondary	0.228	0.270	0.140	-0.002	0.009	-0.278	-0.088
Tertiary	-0.092	0.067	-0.079	0.058	-0.051	0.103	-0.005
Aromatic/enamine	-0.013	0.094	0.034	-0.153	-0.254	0.113	0.073
Primary_1	-0.135	-0.110	0.129	-0.115	0.083	-0.049	-0.166
Secondary_1	0.082	-0.007	0.018	-0.092	-0.012	0.008	-0.237
Tertiary_1	0.165	0.067	0.024	-0.071	-0.024	-0.096	0.017
Vinyl alcohol	-0.033	0.040	-0.147	-0.102	0.030	0.070	0.027
Phenol	0.343	0.238	0.023	-0.022	0.033	0.167	-0.184
Carboxylic	-0.011	-0.124	-0.037	0.001	-0.004	0.091	0.220

Ketone	0.028	-0.032	-0.118	-0.132	-0.108	0.094	0.094
Thioester	0.069	-0.085	0.099	-0.021	-0.038	-0.005	-0.076
Oxime	-0.100	0.045	0.022	0.021	0.003	0.000	-0.008
Oxazolidinone	0.032	0.025	-0.132	0.023	-0.095	0.219	-0.158
Urea	-0.016	0.039	0.027	-0.041	-0.071	-0.008	0.030
Guanidine	-0.040	-0.049	0.132	0.091	-0.143	0.026	-0.110
Ether	0.049	-0.030	0.017	0.024	-0.042	0.039	-0.026
Sulfonamide	-0.058	-0.026	0.320	0.278	-0.215	0.416	-0.014
Sulfone	-0.349	0.327	0.047	-0.101	-0.104	-0.034	-0.176
N-Oxide	0.119	0.106	0.096	0.061	0.067	0.067	0.038
Thioether	0.022	-0.233	0.132	-0.062	0.137	0.008	-0.114
Fluorine	0.094	0.114	-0.183	0.119	-0.197	-0.005	0.003
Pyridine	0.353	-0.121	-0.356	0.245	-0.226	-0.136	-0.369
Alkyl halide	0.021	0.127	0.227	-0.209	0.184	0.129	-0.216
Aryl halide	-0.123	-0.042	0.048	0.002	-0.110	0.124	-0.125
Alkene	-0.068	0.122	0.129	-0.088	-0.003	0.009	-0.154
Alkylgreater than5 C	-0.123	-0.083	0.038	-0.029	-0.136	0.338	-0.158
Phosphonate	-0.068	0.138	0.130	-0.214	-0.222	-0.092	0.385
Hydrozone	0.039	-0.320	0.278	-0.120	0.195	0.083	-0.072
Other_1	-0.229	0.145	-0.248	0.233	0.429	0.247	-0.069
Phosphate	-0.022	0.093	0.045	-0.024	-0.059	-0.036	-0.072
Carbamate	0.111	0.102	0.068	-0.074	-0.333	0.202	-0.101
Nitro	-0.028	0.103	-0.085	0.027	-0.070	-0.098	-0.118
Nitrate	0.119	0.106	0.096	0.061	0.067	0.067	0.038
Steroid	0.057	0.086	0.080	0.035	-0.087	-0.059	-0.014
Hormone	-0.127	-0.129	-0.036	0.204	-0.110	-0.079	0.143
O-heterocyclic	-0.272	-0.206	-0.261	-0.246	-0.292	-0.183	-0.178
N-heterocyclic	0.224	-0.016	-0.055	-0.100	0.032	0.073	0.210
S-heterocyclic	0.155	0.297	0.002	0.230	0.134	0.015	0.089
Long alkyl	0.012	0.041	-0.051	0.226	-0.100	0.010	0.203
Phenyl ring	0.026	0.005	-0.232	-0.330	0.255	-0.065	-0.118
Erythromycin deriv	-0.100	0.045	0.022	0.021	0.003	0.000	-0.008
Tetracycline	-0.100	0.045	0.022	0.021	0.003	0.000	-0.008
Macrocyclic	0.004	0.072	-0.080	-0.033	0.033	0.105	0.048
Macrolide	0.245	-0.202	-0.101	-0.068	-0.020	0.077	0.071
Benzodiazepine	-0.138	0.355	-0.026	-0.095	-0.044	-0.074	-0.108
Barbiturate	0.018	-0.007	-0.039	0.093	0.048	0.038	-0.273
Water	-0.225	-0.039	0.112	0.413	0.049	-0.275	-0.097
Ethanol	0.093	-0.158	0.343	0.111	-0.109	-0.244	-0.143
HCl	0.117	0.033	0.061	-0.140	-0.060	0.153	-0.027
Na+	-0.046	0.077	0.045	-0.003	-0.036	-0.057	-0.015
Gd3+	-0.068	0.055	-0.100	0.042	-0.142	0.042	0.022

Variable	PC29	PC30	PC31	PC32	PC33	PC34	PC35
Primary	-0.203	-0.005	-0.384	-0.076	0.123	0.042	-0.060
Secondary	-0.089	-0.107	0.095	0.198	0.068	-0.002	0.100
Tertiary	-0.064	0.119	0.050	-0.074	0.092	0.257	-0.159
Aromatic/enamine	-0.249	-0.047	-0.005	-0.151	-0.017	0.089	0.005
Primary_1	0.017	-0.013	-0.212	-0.003	0.154	0.085	-0.065
Secondary_1	0.092	-0.034	0.003	0.135	0.119	-0.102	-0.119
Tertiary_1	-0.094	-0.136	0.001	-0.030	0.184	0.087	-0.023
Vinyl alcohol	0.009	0.053	0.067	-0.110	0.088	0.217	0.134
Phenol	-0.107	0.138	0.295	-0.201	0.082	-0.246	-0.031
Carboxylic	0.021	0.013	-0.102	-0.132	-0.139	-0.113	0.026
Ketone	0.179	-0.101	-0.056	-0.051	0.474	-0.263	0.058
Thioester	0.007	0.070	-0.058	0.143	0.009	-0.046	0.184
Oxime	0.019	-0.002	0.006	0.002	-0.003	-0.026	0.015
Oxazolidinone	0.090	0.051	0.108	0.130	-0.038	0.070	-0.061
Urea	-0.034	0.132	0.048	0.137	0.007	-0.005	0.049
Guanidine	-0.036	-0.318	-0.080	0.119	0.008	0.142	0.337
Ether	0.016	0.071	-0.075	0.086	-0.091	0.013	0.047
Sulfonamide	-0.030	-0.026	-0.005	-0.029	0.078	0.064	-0.121
Sulfone	0.013	-0.030	0.039	0.012	0.097	0.046	-0.026
N-Oxide	0.022	0.008	0.005	0.003	0.017	0.040	-0.058
Thioether	-0.112	0.026	0.086	0.081	0.011	-0.011	-0.106
Fluorine	0.005	0.158	-0.162	-0.068	-0.043	0.130	-0.109
Pyridine	-0.151	0.122	0.038	0.104	-0.075	0.079	0.134
Alkyl halide	-0.171	0.259	-0.194	0.089	-0.220	0.331	0.071
Aryl halide	0.065	0.269	-0.337	-0.157	-0.033	-0.175	0.184
Alkene	-0.065	-0.001	0.121	0.042	0.039	-0.073	0.020

Alkylgreater than5 C	0.284	0.075	0.198	0.183	0.023	-0.016	-0.072
Phosphonate	-0.062	0.388	0.178	0.321	-0.014	-0.040	0.058
Hydrozone	0.034	-0.051	0.052	0.064	0.105	0.095	0.070
Other_1	-0.395	-0.007	-0.160	0.202	0.241	-0.123	0.019
Phosphate	0.166	-0.115	-0.058	-0.089	0.005	0.078	-0.135
Carbamate	-0.003	0.048	-0.174	0.054	0.171	0.034	-0.103
Nitro	-0.051	0.002	0.012	-0.029	-0.088	-0.053	-0.295
Nitrate	0.022	0.008	0.005	0.003	0.017	0.040	-0.058
Steroid	-0.033	-0.177	0.098	0.100	0.018	0.013	-0.227
Hormone	-0.093	-0.293	0.213	0.282	0.066	0.248	-0.098
O-heterocyclic	-0.066	-0.015	-0.121	0.057	0.118	-0.091	0.122
N-heterocyclic	0.001	-0.042	-0.197	0.006	0.298	-0.084	-0.310
S-heterocyclic	0.461	0.002	-0.141	0.164	0.100	-0.002	0.397
Long alkyl	0.085	0.026	-0.331	0.190	-0.009	0.279	-0.159
Phenyl ring	0.366	0.098	-0.015	0.227	0.009	0.227	-0.240
Erythromycin deriv	0.019	-0.002	0.006	0.002	-0.003	-0.026	0.015
Tetracycline	0.019	-0.002	0.006	0.002	-0.003	-0.026	0.015
Macrocyclic	-0.041	0.048	0.066	-0.122	0.044	0.202	0.070
Macrolide	0.004	0.108	-0.097	0.021	-0.064	0.040	-0.084
Benzodiazepine	0.067	-0.109	-0.006	0.007	-0.027	-0.042	-0.043
Barbiturate	0.089	-0.183	-0.101	-0.118	0.021	0.062	-0.053
Water	0.100	0.331	-0.004	0.053	-0.022	-0.256	-0.264
Ethanol	0.110	0.129	-0.009	-0.028	0.249	0.054	-0.022
HCl	-0.055	-0.278	-0.250	0.380	-0.402	-0.377	-0.127
Na+	0.235	-0.201	-0.019	-0.338	-0.319	0.037	-0.074
Gd3+	0.045	-0.007	-0.010	0.086	-0.005	-0.018	0.100

Variable	PC36	PC37	PC38	PC39	PC40	PC41	PC42
Primary	0.095	0.125	-0.283	0.152	-0.094	-0.073	-0.126
Secondary	-0.202	-0.157	0.421	-0.147	-0.122	0.032	-0.006
Tertiary	-0.290	0.226	-0.228	-0.115	-0.272	-0.133	0.193
Aromatic/enamine	-0.074	-0.082	0.161	-0.363	-0.012	-0.174	-0.389
Primary_1	-0.085	0.160	0.070	-0.003	0.158	0.087	-0.149
Secondary_1	0.231	-0.190	-0.019	0.076	-0.266	-0.203	0.253
Tertiary_1	0.011	-0.014	0.038	-0.092	0.603	-0.070	0.228
Vinyl alcohol	0.168	-0.051	-0.021	0.072	-0.085	0.045	-0.127
Phenol	0.121	0.237	-0.127	0.047	-0.001	0.042	0.040
Carboxylic	0.224	0.070	0.075	-0.031	0.425	0.099	-0.051
Ketone	-0.403	0.077	0.087	0.228	0.024	-0.104	-0.165
Thioester	0.190	-0.032	0.190	0.180	-0.055	-0.197	-0.280
Oxime	-0.012	-0.003	0.001	0.028	-0.051	0.019	-0.035
Oxazolidinone	0.061	-0.044	0.113	-0.078	0.035	0.021	0.048
Urea	-0.005	0.006	-0.036	0.064	0.057	-0.129	0.028
Guanidine	-0.036	-0.074	-0.148	0.252	0.019	0.019	0.272
Ether	0.057	-0.019	0.010	0.177	-0.096	0.336	-0.167
Sulfonamide	0.087	0.118	-0.159	0.128	0.103	-0.068	-0.089
Sulfone	-0.068	0.020	-0.027	-0.094	0.099	-0.222	0.029
N-Oxide	-0.080	-0.041	0.011	0.025	0.038	-0.033	-0.038
Thioether	-0.116	-0.100	-0.046	0.145	0.090	-0.062	0.099
Fluorine	-0.131	0.189	0.108	-0.001	0.071	0.191	0.245
Pyridine	-0.103	0.017	-0.113	0.023	0.095	-0.097	-0.120
Alkyl halide	0.068	0.182	0.230	0.051	0.051	0.047	0.083
Aryl halide	-0.149	-0.388	0.032	-0.117	-0.034	-0.022	0.163
Alkene	0.024	0.220	0.022	0.088	-0.131	0.097	-0.174
Alkylgreater than5 C	0.036	-0.088	0.207	-0.066	0.070	0.036	0.047
Phosphonate	0.022	0.000	-0.091	0.128	0.016	0.010	0.075
Hydrozone	-0.142	0.055	-0.131	-0.250	-0.052	0.099	0.027
Other_1	0.143	-0.051	0.099	0.027	0.039	-0.086	-0.061
Phosphate	-0.009	-0.011	-0.010	-0.078	-0.128	0.480	-0.134
Carbamate	0.061	0.059	0.012	0.022	0.031	-0.010	-0.020
Nitro	-0.006	-0.185	0.067	0.370	0.071	0.166	-0.001
Nitrate	-0.080	-0.041	0.011	0.025	0.038	-0.033	-0.038
Steroid	0.333	-0.284	-0.322	-0.282	-0.001	-0.035	-0.111
Hormone	0.026	0.249	0.241	0.051	-0.026	0.022	-0.010
O-heterocyclic	0.311	0.246	0.018	-0.163	-0.069	0.053	0.137
N-heterocyclic	0.212	0.021	0.240	0.004	-0.173	0.099	0.220
S-heterocyclic	0.171	0.234	-0.079	-0.191	-0.009	-0.020	0.019
Long alkyl	-0.053	-0.194	-0.011	-0.003	-0.054	-0.104	-0.022
Phenyl ring	-0.042	-0.004	-0.183	0.101	0.165	-0.142	-0.190
Erythromycin deriv	-0.012	-0.003	0.001	0.028	-0.051	0.019	-0.035

Tetracycline	-0.012	-0.003	0.001	0.028	-0.051	0.019	-0.035
Macrocyclic	0.086	-0.169	0.086	0.062	-0.088	0.080	0.004
Macrolide	0.028	0.081	0.090	-0.167	-0.066	-0.158	-0.153
Benzodiazepine	-0.021	-0.072	-0.038	-0.050	0.187	0.123	0.018
Barbiturate	-0.025	0.001	0.018	-0.058	-0.011	0.089	-0.089
Water	-0.023	0.149	0.080	-0.141	0.039	-0.044	-0.066
Ethanol	0.130	-0.064	0.104	0.030	0.082	-0.016	-0.049
HCl	-0.166	0.163	-0.073	-0.074	-0.020	-0.014	0.009
Na+	0.137	0.144	0.244	0.220	-0.087	-0.437	0.043
Gd3+	0.004	-0.067	-0.083	0.167	0.046	0.100	-0.275

Variable	PC43	PC44	PC45	PC46	PC47	PC48	PC49
Primary	-0.105	-0.053	0.000	-0.003	0.042	-0.000	0.000
Secondary	0.080	-0.172	0.083	0.101	-0.096	0.000	0.000
Tertiary	0.250	0.055	-0.074	0.329	-0.091	0.000	0.000
Aromatic/enamine	-0.228	0.017	-0.243	-0.028	0.034	-0.000	0.000
Primary_1	0.252	-0.160	0.061	0.096	-0.021	0.000	0.000
Secondary_1	-0.486	0.105	-0.156	-0.046	0.017	-0.000	-0.000
Tertiary_1	0.199	0.242	-0.250	-0.023	-0.004	-0.000	-0.000
Vinyl alcohol	-0.049	-0.091	0.034	-0.182	-0.663	0.000	-0.000
Phenol	0.071	0.034	0.019	-0.053	0.026	-0.000	-0.000
Carboxylic	-0.269	-0.147	0.171	0.369	-0.094	0.000	0.000
Ketone	-0.091	-0.073	0.010	0.108	0.000	-0.000	0.000
Thioester	0.143	0.356	-0.037	0.194	-0.011	-0.000	-0.000
Oxime	-0.022	-0.034	0.014	0.004	0.001	0.003	0.064
Oxazolidinone	0.050	-0.031	-0.011	0.021	-0.021	0.000	-0.000
Urea	0.001	-0.121	-0.013	0.023	-0.002	-0.016	0.010
Guanidine	0.038	-0.024	0.266	0.077	-0.050	0.000	0.000
Ether	0.057	-0.310	-0.443	0.183	-0.026	-0.000	0.000
Sulfonamide	-0.113	0.022	-0.001	-0.048	0.025	-0.000	-0.000
Sulfone	-0.062	0.110	0.221	-0.037	0.007	0.000	0.000
N-Oxide	0.001	-0.007	0.036	0.007	-0.022	0.707	0.000
Thioether	-0.005	-0.239	0.021	-0.129	0.003	0.008	-0.012
Fluorine	-0.149	-0.200	0.022	-0.197	0.013	-0.000	0.000
Pyridine	-0.083	-0.062	-0.007	0.080	0.002	-0.000	0.000
Alkyl halide	-0.013	0.117	-0.020	-0.054	0.005	-0.000	-0.000
Aryl halide	0.143	-0.019	0.011	0.011	-0.063	0.000	0.000
Alkene	0.030	-0.192	0.144	0.081	-0.007	0.000	0.000
Alkylgreater than5 C	0.094	-0.056	0.001	0.051	-0.041	0.000	-0.000
Phosphonate	-0.019	0.003	0.058	0.052	-0.012	0.010	-0.006
Hydrozone	-0.113	-0.135	-0.050	-0.181	0.056	-0.006	0.010
Other_1	-0.005	-0.071	0.021	0.016	-0.009	0.000	0.000
Phosphate	-0.094	0.398	0.122	0.093	-0.020	-0.002	-0.003
Carbamate	0.027	-0.079	-0.015	0.043	-0.010	-0.000	0.000
Nitro	0.039	0.127	0.162	0.091	-0.067	-0.006	0.010
Nitrate	0.001	-0.007	0.036	0.007	-0.022	-0.707	-0.000
Steroid	0.296	-0.238	0.082	0.144	-0.045	0.000	0.000
Hormone	-0.102	0.022	-0.027	-0.041	0.013	-0.000	-0.000
O-heterocyclic	0.085	-0.029	-0.027	0.005	0.087	-0.000	0.000
N-heterocyclic	0.178	-0.022	-0.014	-0.106	-0.020	0.000	-0.000
S-heterocyclic	-0.002	-0.020	-0.017	0.020	0.043	-0.000	0.000
Long alkyl	-0.041	-0.018	-0.030	-0.020	-0.014	0.000	-0.000
Phenyl ring	-0.054	-0.011	-0.054	0.073	0.067	-0.000	0.000
Erythromycin deriv	-0.022	-0.034	0.014	0.004	0.001	-0.002	-0.737
Tetracycline	-0.022	-0.034	0.014	0.004	0.001	-0.002	0.673
Macrocyclic	-0.011	-0.126	0.115	0.169	0.676	-0.000	0.000
Macrolide	0.026	-0.024	0.610	-0.157	0.020	0.000	0.000
Benzodiazepine	-0.206	-0.266	0.040	-0.166	0.033	-0.000	0.000
Barbiturate	-0.033	0.062	0.006	0.036	-0.004	0.014	-0.009
Water	-0.010	0.022	-0.030	-0.042	0.003	-0.000	-0.000
Ethanol	-0.055	-0.055	0.037	-0.041	0.010	0.000	-0.000
HCl	0.027	0.044	-0.030	0.004	-0.019	0.000	-0.000
Na+	0.166	-0.238	-0.101	-0.041	0.008	0.000	0.000
Gd3+	0.307	0.074	-0.058	-0.590	0.184	-0.000	0.000

Variable	PC50	PC51	PC52
Primary	0.000	0.000	-0.000
Secondary	0.000	0.000	0.000
Tertiary	0.000	-0.000	0.000
Aromatic/enamine	0.000	-0.000	-0.000

Primary_1	-0.000	-0.000	0.000
Secondary_1	0.000	0.000	-0.000
Tertiary_1	-0.000	-0.000	0.000
Vinyl alcohol	0.000	-0.000	0.000
Phenol	-0.000	-0.000	-0.000
Carboxylic	0.000	0.000	0.000
Ketone	0.000	0.000	0.000
Thioester	0.000	-0.000	0.000
Oxime	-0.066	0.811	0.009
Oxazolidinone	-0.000	-0.000	0.000
Urea	0.370	0.035	-0.530
Guanidine	0.000	0.000	0.000
Ether	-0.000	-0.000	-0.000
Sulfonamide	-0.000	-0.000	-0.000
Sulfone	0.000	0.000	0.000
N-Oxide	0.004	-0.002	-0.018
Thioether	0.462	0.034	0.401
Fluorine	-0.000	0.000	-0.000
Pyridine	0.000	0.000	0.000
Alkyl halide	-0.000	-0.000	0.000
Aryl halide	0.000	-0.000	0.000
Alkene	0.000	0.000	0.000
Alkylgreater than5 C	0.000	-0.000	0.000
Phosphonate	-0.234	-0.022	0.335
Hydrozone	-0.379	-0.028	-0.330
Other_1	-0.000	0.000	0.000
Phosphate	0.436	0.036	-0.004
Carbamate	-0.000	-0.000	0.000
Nitro	-0.379	-0.028	-0.330
Nitrate	-0.004	0.002	0.018
Steroid	-0.000	0.000	0.000
Hormone	-0.000	-0.000	-0.000
O-heterocyclic	-0.000	0.000	-0.000
N-heterocyclic	-0.000	-0.000	-0.000
S-heterocyclic	-0.000	0.000	-0.000
Long alkyl	0.000	-0.000	-0.000
Phenyl ring	-0.000	-0.000	0.000
Erythromycin deriv	0.024	-0.350	-0.020
Tetracycline	0.042	-0.461	0.012
Macrocyclic	-0.000	0.000	-0.000
Macrolide	0.000	0.000	-0.000
Benzodiazepine	-0.000	0.000	-0.000
Barbiturate	-0.333	-0.031	0.477
Water	-0.000	-0.000	-0.000
Ethanol	-0.000	0.000	-0.000
HCl	-0.000	-0.000	0.000
Na+	-0.000	-0.000	-0.000
Gd3+	-0.000	-0.000	-0.000

Figure I PCA data for analysis of chemical data given in the case studies.